Proceedings of the Design Society, Volume 5: ICED25
https://doi.org/10.1017/pds.2025.10340

Uncovering the limits of visual-language models in engineering knowledge representation

Marco Consoloni [1,2,✉], **Vito Giordano** [1,2], **Federico Andrea Galatolo** [1], **Mario Giovanni Cosimo Antonio Cimino** [1] and **Gualtiero Fantoni** [1,2]

[1] *University of Pisa, Italy,* [2] *Business Engineering for Data Science (B4DS) research group, Italy*

✉ marco.consoloni@phd.unipi.it

ABSTRACT: Visual-Language (VL) models offer potential for advancing Engineering Design (ED) by integrating text and visuals from technical documents. We review VL applications across ED phases, highlighting three key challenges: (i) understanding how functional and structural information is complementarily expressed by text and images, (ii) creating large-scale multimodal design datasets and (iii) improving VL models' ability to represent ED knowledge. A dataset of 1.5 million text-image pairs and an evaluation dataset for cross-modal information retrieval were developed using patents. By Fine-tuning and testing the CLIP base model on these datasets, we identified significant limitations in VL models' capacity to capture fine-grained technical details required for precision-driven ED tasks. Based on these findings, we propose future research directions to advance VL models for ED applications.

KEYWORDS: artificial intelligence, machine learning, design engineering, visual-language models, patents

1. Introduction

When we interact with the world surrounding us, we see objects, hear sounds, fell textures and smell odours. This process involves multiple modalities. A modality refers to the medium through which an object exists or is experienced (BaltruŠaitis et al., 2018). Common modalities include natural language text, visuals signals (e.g., images or videos), and audio signals.

Throughout the process of Engineering Design (ED), ED knowledge is communicated and evolves across different modalities. In the early stages of ED, abstract representations such as free-hand sketches and handwritten textual descriptions enable designers to effectively communicate their ideas, quickly modify their designs and explore the design space. In later stages, high-fidelity representations such as detailed technical drawings and written design specifications are used to, support design communication, optimization and manufacturing. Many design documents are multimodal, for example, product design specifications combine CAD models with textual descriptions, patents use images and text to describe patent devices, and assembly manuals pair visuals with textual instructions.

While multiple modalities are commonly used in design practice, most Artificial Intelligence (AI) applications supporting design research rely on unimodal approaches (Song et al., 2024). In order for AI to make progress in the field of ED, it needs to be able to interpret and capture multimodal information from technical documents. Multimodal AI focuses on creating models that can process and relate information from different modalities (BaltruŠaitis et al., 2018). These models offer great potential for creating systems that can leverage images and text in technical documents, providing a deeper understanding of the design space and supporting data-driven design applications. Among existing multimodal models, Visual-Language (VL) models are AI models that use visual information (images or video) with textual information (natural language text). In this work we focus only on text and image modalities, reviewing current applications of VL models in the context of ED. Moreover, we develop a

https://doi.org/10.1017/pds.2025.10340 Published online by Cambridge University Press

VL model specifically trained on patent documents for retrieving patents using both patent text and drawings.

2. Visual-language models for engineering design

We performed a literature review of VL models within the context of ED applications. VL models can be divided into foundational models and fine-tuned models. Foundational models are trained on large, general datasets, so they can be applied across a wide range of tasks. Examples of foundational VL models are OpenAI's GPT-4V (vision), DALL-E and Flamingo. These models are not specifically trained on design data. Fine-tuned models are created by adapting (finetuning) foundational models using domain-specific dataset, improving their performance on specialized tasks. Figure 1 maps key ED tasks across the ED phases: problem definition, conceptual, embodiment and detailed design. These ED tasks are further classified based on whether they have been addressed using foundational VL models, fine-tuned VL models, or remain as gaps in current research.

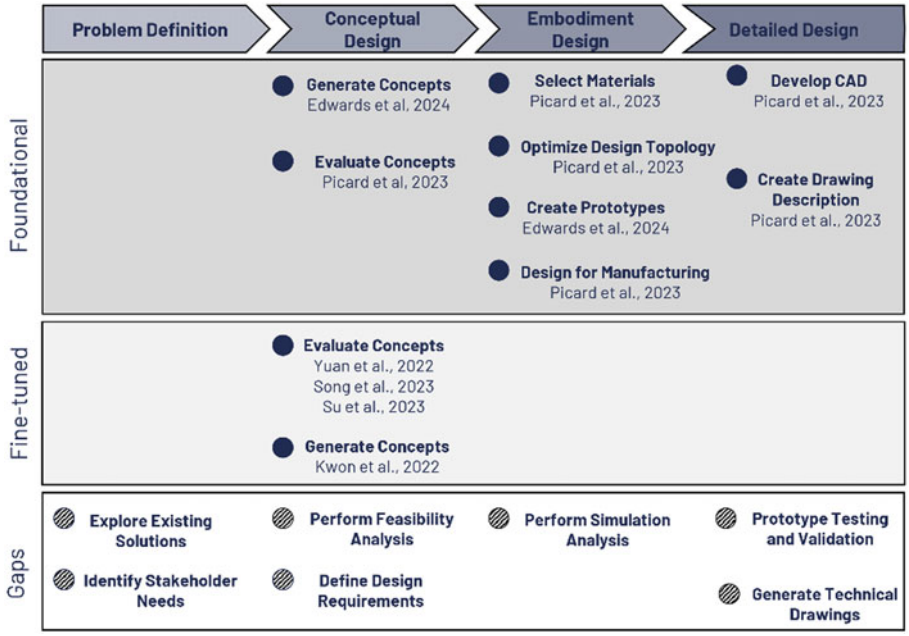

Figure 1. Visual-language models for engineering design tasks

Figure 1 shows that foundational VL models have been used for ED tasks across the conceptual, embodiment and detailed design phases. In the conceptual design phase, Edwards et al. (2024) used OpenAI GPT-4V (Vision) and DALL-E 3 to generate design concepts using sketches and textual descriptions. Picard et al. (2023) used ChatGPT-4V and Llava 1.6 34B to assess design similarity of design sketches for concept selection. In the embodiment and detailed design phases, Picard et al. (2023) used text and image prompts to perform tasks such as selecting materials, optimizing design topology, generating drawing descriptions, and developing CAD models. However, the effectiveness of these foundational VL models for precision-driven ED tasks remains limited. In fact, these models struggle to "understand" technical details, and their answers are never fully accurate (Edwards et al., 2024). For instance, their current capabilities are not yet sufficient for interpreting complex technical drawings or producing viable CAD outputs (Picard et al., 2023). This limitation stem from the fact that foundational VL models have not been specifically trained on design data and their underlying design knowledge remains limited (Song et al., 2024). As a result, they struggle to capture fine-grained design concepts. To address this issue, several studies have fine-tuned foundational VL models using design-specific data. Yuan et al. (2022) developed a VL model to evaluate design concepts using product reviews with images and text. Similarly, Su et al. (2023) used 2,571 instances of online textual and visual information to develop a VL model for evaluating vehicle designs. Song et al. (2023) fine-tuned a VL model to assess creativity of 1,086 freehand sketches with handwritten descriptions. Kwon et al. (2022), using a dataset comprising 2D snapshots of 26,671 3D objects, finetuned a VL model capable of retrieving design

https://doi.org/10.1017/pds.2025.10340 Published online by Cambridge University Press

stimuli by combining images with keywords describing component features in order to generate new design concepts.

As shown in Figure 1, fine-tuned VL models, have been used only in the conceptual design phase for design concept evaluation and generation. Furthermore, all previous approaches relied on small-scale design datasets for fine-tuning and employed simple images such as sketches rather than complex technical drawings. Moreover, the applications of fine-tuned VL models to other ED tasks, such as exploring existing solutions, identifying stakeholder needs, and generating technical drawings, remains unaddressed in the current literature due to the following main limitations.

(1) Multimodal Design Dataset Creation: the creation of large multimodal data for VL model finetuning is challenging because large accessible design datasets like Pinterest and Fusion 360 Gallery contain single modality data and manual labelling design data is time-consuming and resource-intensive (Song et al., 2024, Jin et al., 2024). Moreover, creating multimodal dataset involve aligning textual descriptions with corresponding images. This imply finding semantic relationships and correspondences between text and visuals. For example, given a technical drawing and a caption we must find areas of the drawing corresponding to the caption's words or phrases. This is challenging because ED concepts like functions, components, unit of measurements and spatial orientations are often not explicitly represented in either the text or the visuals. Moreover, the relationships between modalities are often subjective (Baltrusaitis et al., 2018). A single image can be described by multiple texts and vice versa, and a "correct description" of a technical drawing may not exist. For instance, Figure 2 demonstrates alternative textual and visual representations of a bearing-shaft assembly, where combinations of text-images pairs (T1-I1; T2-I2) convey the same information while complementing each other. As a results, there is a lack of high-quality multimodal datasets for ED applications. (Kwon et al., 2022; Picard et al., 2023; Song et al., 2023; Song et al., 2024; Li et al., 2023; Consoloni et al., 2024, Jin et al., 2024).

(2) Effective Representations of Design Knowledge: VL models are required to learn how to represent and join information from text and visuals in a way that exploits the complementarity and redundancy of the two modalities (Baltrusaitis et al., 2018). In the context of ED, this is challenging and still far beyond an ideal level (Pan et al., 2024), as it requires domain-specific understanding of how structural and functional design features are expressed across text and images (Kwon et al., 2022; Picard et al., 2023; Song et al., 2023; Song et al., 2024). For example, Figure 2 illustrates how functional and structural information can be distributed between visual and textual descriptions when explaining a bearing-shaft assembly. Text segment T1 encodes functional information through phrases like *"shaft rotates"* and *"withstands radial forces"* while the corresponding image I1 conveys structural details, such as the shaft having a taper on the left-hand side using arrows to indicate dimensions. Conversely, an alternate case can occur, where text segment T2 describes structural information using terms like *"coaxially mounted"* and *"20 mm taper"*, whereas the corresponding image I2 represents functional information through visual elements such as arrows indicating rotation and force resistance.

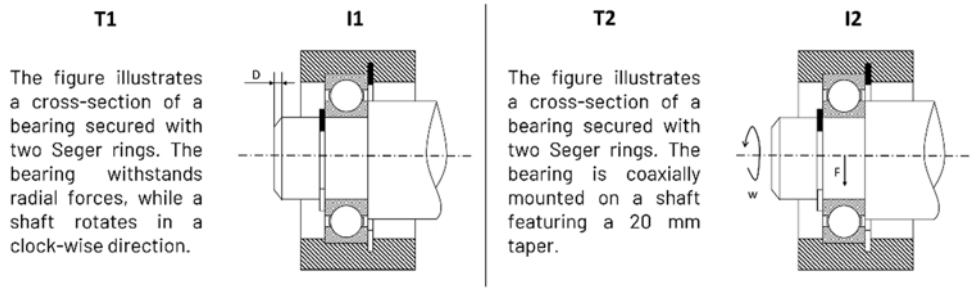

Figure 2. Two alternative uses of text and images to express functional and structural ED concepts

3. The task: patent citations retrieval

In this work we develop a fine-tuned VL model for exploring existing solutions using a large-scale multimodal patent dataset consisting of 1.5M text-image pairs This task falls under the problem definition phase, and it has not been addressed with VL models in previous research.

In the field of ED, exploring existing solutions is a critical ED task. Patent databases are valuable repositories of technical knowledge and serve as a key source of information on prior art. When

https://doi.org/10.1017/pds.2025.10340 Published online by Cambridge University Press

designing a new product, conducting effective prior art searches using patents enables designers to identify technical challenges, evaluate gaps in current solutions, and access a diverse array of design stimuli. However, this task poses significant challenges for engineers because it requires (1) capturing fine-grained technical similarities and differences between the proposed design and existing patented devices and, (2) a comprehensive understanding of patent text and drawings. Additionally, commercial patent search platforms rely solely on textual keyword searches, lacking image-based querying. This limitation affects prior art retrieval due to language ambiguity and terminology variations, often requiring iterative query expansion to accurately capture complex technological domains. For these reasons, VL models offer a promising solution not only to use both patent text and drawings to search patent database but also for facilitating cross-source retrieval by connecting patents with CAD models and online product reviews.

In this work, we finetuned the foundational VL model, developed by OpenAI, known as Contrastive Language-Image Pre-training (CLIP) model (Radford et al., 2021), to perform prior art search on patent databases using both patent text and drawings. Specifically, we tested its performance on an extremely knowledge-intensive task: retrieving patent citations made by examiners. Patent citations are references to existing patents which are added by patent examiners during the examination process of patent applications. When an existing patent is cited, it indicates that examiners have assessed it as technically relevant to a pending application, potentially challenging the application's patentability. By choosing this task, we aimed to evaluate our fine-tune VL model on a task where precision, high-level reasoning and domain-specific knowledge are critical.

4. Methodology

The methodology proposed in this work is composed of three phases: (1) *Data Collection and Preprocessing* involving the retrieval and processing of patent documents to create a large-scale multimodal patent dataset (2) *Model Fine Tuning*, where we fine-tuned the CLIP model on our dataset; and (3) *Model Testing*, evaluating the performance of the fine-tuned CLIP model in retrieving patent citations.

4.1. Data collection and preprocessing

We collected all utility patents granted between 2020 and 2024 from the United States Patent and Trademark Office (USPTO) *Bulk Data Storage System (https://bulkdata.uspto.gov/)*. Following the approach proposed by Consoloni et al. (2024), we extracted: 1) the front image from the cover page, which is considered as representative of the invention, and 2) the first claim, which outlines the main features and elements of the patented device for which protection is sought. Each front image was aligned with its corresponding first claim, and patents missing either were excluded from the dataset. The final dataset comprises 1,383,944 text-image pairs, totalling 68GB (8GB text, 60GB images). This phase, which involved downloading and decompressing .tar archives, extracting images and textual data from XML files, required approximately 100 hours (4 days) to complete.

4.2. Model fine tuning

We fine-tuned the CLIP model *(https://huggingface.co/openai/clip-vit-base-patch32)* on our dataset. CLIP is a foundational multimodal model pre-trained on a large-scale generic dataset of text-image pairs. Its training objective is to match texts with corresponding images (Radford et al., 2021). Specifically, the model encodes input images and their corresponding textual descriptions into high-dimensional numerical vectors, known as embeddings. These embeddings are mapped into a shared vector space, known as embedding space, where both images and text are represented as points. If an image and a text have similar meanings, their points are placed close to each other within the embedding space. If they are semantically unrelated, their points are placed far apart. This shared embedding space enables the use of distance measures, such as cosine similarity, to retrieve relevant texts based on an input image, or vice versa. As a result, the model is well-suited for cross-modal information retrieval tasks (Galatolo et al., 2021).

For this work, we divided our dataset into training (1,107,155 pairs), validation (138,394 pairs), and test sets (138,395 pairs) following an 80:10:10 split. Then, we fine-tuned the CLIP base model using a contrastive loss which force the model to match first claims with their corresponding front images. After fine-tuning, our model generates 512-dimensional embeddings for first claims and front images within a

https://doi.org/10.1017/pds.2025.10340 Published online by Cambridge University Press

shared embeddings space. First claims were truncated to the first 77 tokens to comply with the CLIP text encoder's input limit. These tokens do not necessarily correspond to words but can include sub words and punctuation. The fine-tuning process involved training the CLIP base model for 100 epochs with a batch size of 256. The training was conducted on an ARM Neoverse-N1 CPU (256 cores) with 1 TB of RAM and an NVIDIA A100 GPU (PCIe), completing in 3 days.

4.3. Model testing: patent citation retrieval

4.3.1. Test dataset

We evaluated our fine-tuned model in retrieving patents cited by examiners. To create the evaluation dataset, we used web scraping to collect first claims and front images for all US patents in IPC classes A42B3/00, A62B18/00, and H02K19/00 from *Google Patents* (*https://patents.google.com/*). Class A42B3/00 contains patents of "helmets or other protective head coverings", class A62B18/00 includes patents of "breathing masks", and H02K19/00 contains patents of "synchronous motors or generators". These classes were selected to evaluate the model's ability in retrieving and distinguishes patent citations across closely related technological domains (A42B3/00 and A62B18/00) and significantly different ones (H02K19/00). For each successfully scraped patent (citing patent), we randomly selected five patents cited by examiners (cited patents) and scraped their corresponding first claims and front images. Table 1 summarizes the number of citing and cited patents successfully scraped for each IPC class.

Table 1. Experimental dataset for patent citation retrieval task

IPC class	Description	N. of Citing Patents (%)	N. of Cited Patent (%)
A42B3/00	Helmets or other protective head coverings.	143 (61.64)	658 (63.67)
A62B18/00	Breathing masks.	63 (27.16)	263 (25.41)
H02K19/00	Synchronous motors or generators.	26 (11.21)	114 (11.01)
Total		232 (100.00)	1,035 (100.00)

4.3.2. Retrieval workflow

To perform patent citation retrieval, as shows in Figure 3, we used our fine-tuned CLIP model to generate embeddings for first claims and front images of both citing and cited patents. Next, we used *qdrant* (*https://qdrant.tech/*), an open-source vector database and similarity search engine designed to handle high-dimensional vectors, to organize embeddings into three separate collections: (1) **text collection** which contains text embeddings of first claims; (2) **image collection** which contains image embeddings of front images, and (3) **joint collection** which contains embeddings created by summing the embeddings of first claims and their corresponding front image embeddings.

Then, we developed three multimodal strategies to retrieved cited patents. These retrieval strategies utilize cosine similarity to measure the closeness between text and image embeddings within the embeddings space created by our fine-tuned CLIP model. (1) **Text-vs-image retrieval:** given the first claim embedding of a citing patent (text), the system retrieves the top-30 cited patents with the closest front image embeddings (image) in the embedding space; (2) **Image-vs-text retrieval:** given the front image embedding of a citing patent (image), the system retrieves the top-30 cited patents with the closest first claim embeddings (text) in the embedding space; (3) **Joint retrieval:** given the joint embedding of a citing patent (text + image), the system retrieves the top-30 cited patents with the closest joint embeddings in the embedding space.

The text-vs-image (image-vs-text) retrieval strategy tests the model's ability to produce embeddings of first claims (front images) capable of retrieving semantically related front images (first claims). These tasks aim to evaluate the model's capability to perform cross-modal information retrieval (i.e., transfer ED knowledge between text and image modality). In contrast, the joint retrieval strategy tests the model's ability to synthesize ED knowledge into a unified representation.

4.3.3. Retrieval performance evaluation

To evaluate the performance of our fine-tuned CLIP model across the retrieval strategies, we calculated precision, recall, and F1-score at k, where k represents the number of top-ranked results considered (e.g.,

https://doi.org/10.1017/pds.2025.10340 Published online by Cambridge University Press

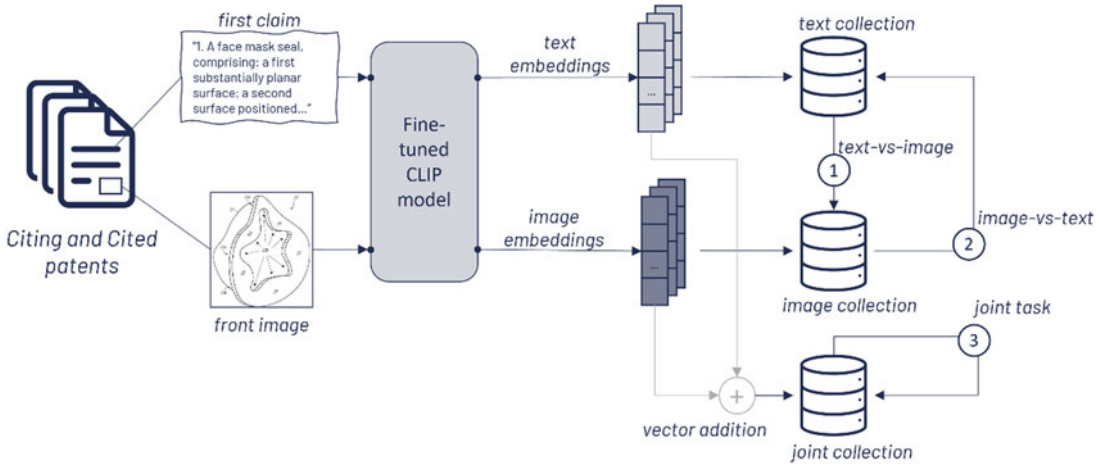

Figure 3. Retrieval Workflow

top 1, top 5, top 10). Specifically, for a given citing patent, these metrics assess how effectively the model retrieves its cited patents:

Precision@k measures the number of cited patents within the top k retrieved patents. Example: if k = 10 and the system retrieved 4 cited patents, the precision at k is 4/10 = 0.4. This metric ranges from 0 to 1 and measures how accurate the system is in retrieving cited patents within the top k results. Ideally, it should be 1; **Recall@k** measure the proportion of cited patents within the top k retrieved patents. Example: if there are 5 cited patents in total, and the system retrieves 4 cited patents within the top k = 10, the recall at k is 4/5 = 0.8. This metrics ranges from 0 to 1 and measures how comprehensive the system is in retrieving cited patents within the top k results. Ideally, it should be 1; **F1-score@k** combines precision and recall at k into a single value by taking their harmonic mean. Example: if Precision@k = 0.4 and Recall@k = 0.8, the F1score at k is 2*(0.4*0.8) / (0.4 + 0.8) = 0.53. This score ranges from 0 to 1 and balances the trade-off between precision and recall.

To evaluate the overall performance of our fine-tuned CLIP model across all citing patents, we compute the following averages: 1) **Avg. P@k:** the mean of Precision@k across all citing patents; 2) **Avg. R@k:** the mean of Recall@k across all citing patents, and 3) **Avg. F1@k:** the mean of F1-score@k across all citing patents.

5. Results and discussions

Table 2 presents the performance results for each retrieval task across k values of 1, 3, and 5, comparing the CLIP base model and our fine-tuned model. Bold numbers indicate the best-performing model for each column. The baseline model achieves its highest F1-score of **0.081** for k=5 on the joint task, indicating that, on average, it is not capable to accurately retrieve the 5 cited patents corresponding to each citing patent. In contrast, the fine-tuned model shows slight but consistent improvements across all tasks and k values, reaching a maximum F1-score of **0.09** at k=5 on the joint task. This indicates the contribution of our fine-tuning to very modest performance gains. However, the overall performance of both VL models remains insufficient for effective multimodal patent citation retrieval as their results are significantly below the optimal value of 1. This indicate that, our fine-tuned CLIP model produces embeddings that fails to capture technical details required for accurate patent citation retrieval within top-5 results.

To analyse the overall performance of our fine-tuned CLIP model, Figure 4 presents the plots of *Avg. P@k*, *Avg. R@k*, and *Avg. F1@k* metrics across k values up to 30 for the text-vs-image, image-vs-text and joint tasks. In all plots, as k increases, *Avg. R@k* increases because retrieving more patents naturally captures more cited patents. Conversely, *Avg. P@k* decreases with higher k because the inclusion of additional retrieved patents introduces more non-cited patents, reducing precision. Given that each citing patent has on average 5 cited patents, as expected, *Avg. F1@k* peaks around k=5 for all retrieval tasks. In fact, beyond k=5, while *Avg. R@k* continues to increase, *Avg. P@k* drops as more non-cited patents are retrieved, causing *Avg. F1@k* to decline.

Figure 4 indicates also that the joint retrieval task demonstrates relatively better performance compared to the text-vs-image and image-vs-text tasks, as reflected by higher F1-scores across all k values.

https://doi.org/10.1017/pds.2025.10340 Published online by Cambridge University Press

Specifically, F1@k for the joint retrieval strategy lies within the range [0.05, 0.1], whereas for the other tasks, it remains within [0.0, 0.05]. This highlight that our fine-tuned CLIP model produces embeddings which do not enable cross-modal retrieval of patent citations (i.e., the model has limited ability to transfer ED knowledge between text and image modality) and, the joint embeddings demonstrate a slightly improved capacity to capture ED knowledge.

Table 2. Results of patent citation retrieval tasks for CLIP base and fine-tuned model

k	retrieval task	Avg. P@k base	Avg. P@k fine-tuned	Avg. R@k base	Avg. R@k fine-tuned	Avg. F1@k base	Avg. F1@k fine-tuned
1	text-vs-image	0.039	0.039	0.009	0.008	0.014	0.014
	image-vs-text	0.043	0.039	0.010	0.011	0.016	0.017
	joint	**0.125**	0.151	0.031	0.038	0.048	0.058
3	text-vs-image	0.032	0.027	0.021	0.018	0.025	0.021
	image-vs-text	0.029	0.030	0.019	0.027	0.023	0.026
	joint	0.098	**0.111**	0.066	0.082	0.077	0.090
5	text-vs-image	0.024	0.024	0.026	0.026	0.025	0.025
	image-vs-text	0.024	0.033	0.028	0.045	0.025	0.036
	joint	0.078	0.085	**0.090**	**0.102**	**0.081**	**0.090**

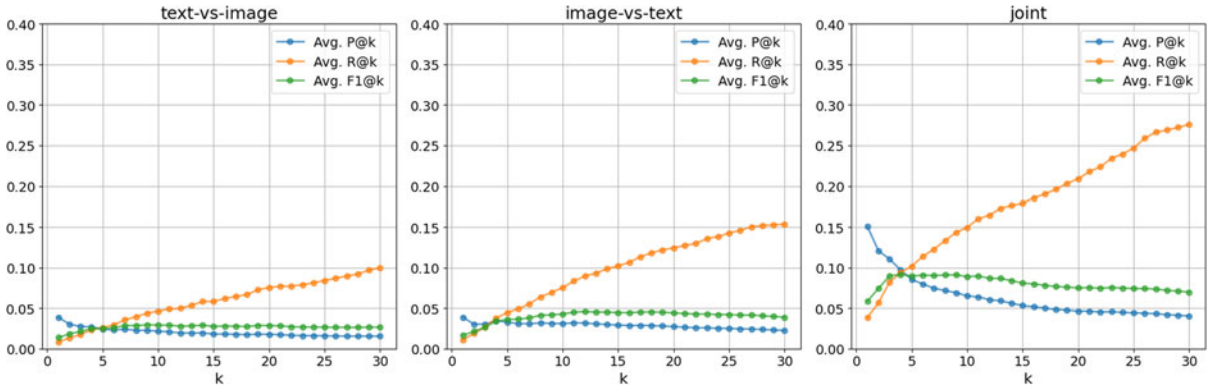

Figure 4. Performance of retrieval strategies: Avg. P@k, R@k and F1@k for k values 1-30

Previous results show that our fine-tuned CLIP model is unable to retrieve the cited patents with performance comparable to that of patent examiners. However, we evaluated the model's capability to distinguish between patents from different IPC classes (i.e., technological domains) during retrieval. For example, given a citing patent which falls under the class A42B3, which relates to "helmets or other protective head coverings," the model should avoid retrieving patents from unrelated classes, such as H02K19, which relates to "synchronous motors or generators". Ideally, for a citing patent belonging to a specific IPC class, we expect the top 30 retrieved patents to belong to the same IPC class, even if they are not the exact cited patents. This expectation underscores the model's ability to generate embeddings that capture general technological domain information, ensuring that retrieved patents remain at least relevant within the broader technological context of the citing patent.

Figure 5 shows the IPC class distribution among the top 30 retrieved patents for citing patents belonging to A42B3, A62B18, and H02K19 IPC classes. It helps to assess the model's ability to maintain domain consistency by prioritizing patents from the same technological class. For example, in the case of class A42B3, 92% of retrieved patents belongs to the same class (A42B3), 7.22% are from A62B18, and 0.77% belong to class H02K19. For all IPC classes, the majority of the top-30 retrieved patents belong to the same IPC class as the citing patents. This shows that the model produces embeddings of first claims and images that can effectively distinguish between IPC classes during retrieval. However, for citing patents in A62B18, 19.21% of top-30 retrieved patents belong to class A42B3. This suggests that the model faces challenges in achieving fine-grained differentiation between closely related domains, such as A42B3 (helmets) and A62B18 (breathing masks). In contrast, it successfully distinguishes these classes from the significantly different technological domain of H02K19 (electric motors).

https://doi.org/10.1017/pds.2025.10340 Published online by Cambridge University Press

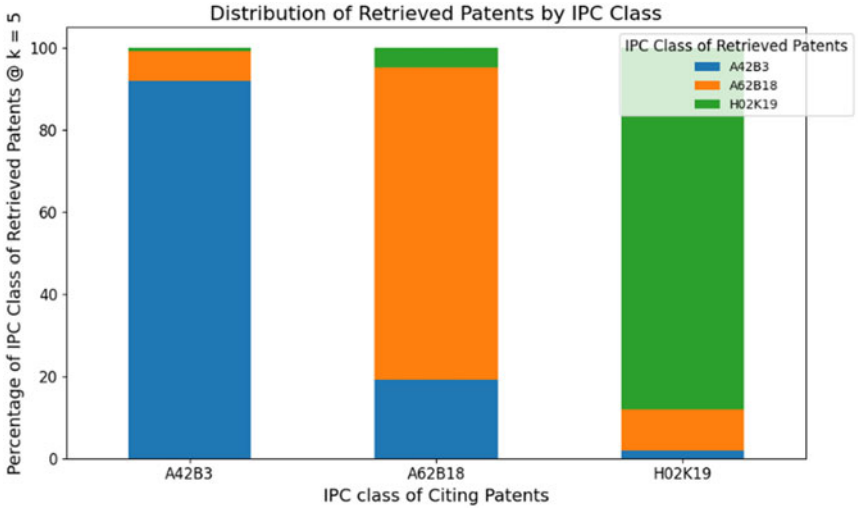

Figure 5. Distribution of IPC classes of retrieved patents for each citing patent

This indicates that despite the use of a large dataset for fine-tuning (1,5M text-image pairs), our model has primarily learned general ED knowledge sufficient for classify patents by IPC class. Notably, in literature patent retrieval has been done with higher accuracy using only text. A relevant example is the study by Siddharth et al. (2022), which utilizes text-only embeddings for patent retrieval in ED applications. This rises a critical concern about multimodality: VL models are resource-intensive models in terms of energy, computing infrastructures and training time. Without effective multimodal representations to fully capture ED knowledge from both text and images, these models risk of being nothing more than oversized, inefficient classifiers.

6. Limitations and future research directions

This section examines the limitations of our methodology and proposes potential improvements for future research. These limitations are categorized into three layers: (1) Data, (2) Model, and (3) Testing, highlighting the specific areas where challenges occur.

The primary limitations related to **(1) Data** are as follows. **(1.1)** one limitation is using only the first claim as input, which relies heavily on technical-legal jargon and captures only surface-level details of inventions. Future research will incorporate additional patent sections, such as abstracts and detailed descriptions to provide a more comprehensive representation of technical details. **(1.2)** Patent first claims describe both functional and structural aspects of a device. This approach uses first claims as input without distinguishing these aspects. Classifying sentences into functional and structural categories using NLP methods could enable the generation of distinct functional and structural descriptions for patent images, enabling functional-structural indexing and retrieval of patent data (Song et al. 2023; Kwon et al. 2022).

The primary limitations related to **(2) Model** are as follows. **(2.1)** One key issue is that CLIP's textual encoder truncates first claims to the initial 77 tokens, potentially omitting critical information required for patent citations retrieval. More precisely, first claims contain an average of 243 tokens (with a standard deviation of 152). As a result, the CLIP model discards approximately 70% of the information in first claims. To address this issue, as suggested by Lo et al. (2024), we plan to use LLMs to summarize functional and structural information of first claims using a prompt template. Additionally, to bypass CLIP's text input limit, a hierarchical encoding approach could be employed. Rather than encoding the entire first claim at once, the text can be divided into smaller segments, each independently encoded to capture local semantic representations. Then, these individual representations can be aggregated into a single global embedding that capture the overall meaning of the first claim. **(2.2)** We implemented an early fusion approach, summing text and image embeddings element-wise to create a single joint representation. This is limiting because it treats both modalities equally and does not capture complex relationships between them. This could be improved by adopting more advanced multimodal fusion techniques, such as attention-based methods, which can capture deeper inter-modality relationships (Song et al., 2023; Li et al., 2023). **(2.3)** Pretrained VL models, such as our finetuned CLIP, operate as

https://doi.org/10.1017/pds.2025.10340 Published online by Cambridge University Press

"black boxes" making it difficult to understand how they process images and text to produce outputs (BaltruŠaitis et al., 2018). This lack of transparency limits the explainability and complicates the analysis of specific failure cases (Song et al., 2024). For instance, it remains unclear which concepts of ED knowledge are represented in our embeddings, and which features of visual and textual inputs our model attends to (prioritize) when constructing the embedding space. As a result, quantitatively determining why some cited patents are retrieved while others are not remain extremely challenging and beyond the scope of this work. To address this, we plan to explore our embedding space using dimensionality reduction techniques such as PCA and t-SNE projection, which may provide macro-level insights into its structure (Lo et al., 2024). Additionally, we also aim to use NLP techniques to identify ED concepts, including functions, components, material and spatial arrangements within first claims. By doing so, we aim to use these identified concepts as explanatory elements to interpret and debug misleading similarities in model's outputs, facilitating the validation of model predictions against established ED knowledge (Song et al., 2023; Kwon et al., 2022). This approach will also contribute to refining the cosine similarity metric by re-weighting embeddings based on the importance of ED concepts, ensuring that similarity computations prioritize technical terms over common words.

The primary limitations related to **(3) Testing** are as follows. **(3.1)** Our methodology does not currently benchmark our VL model with a text-only model for patent citations retrieval. In future studies we plan to use an LLM to generate text embeddings of first claims and use them for our patent citations retrieval task. In fact, when working with VL models, evaluating whether the additional complexities of image processing and text-image alignment are justified by performance gains is crucial for advancing research in this field. **(3.2)** The CLIP base model is pre-trained on coloured RGB images, primarily composed of real-world photographs (natural images). Therefore, it may struggle to process black-and-white schematic images found in patents, as these differ significantly in style, structure, and complexity from the natural images used during its pre-training. To evaluate this pretraining biases, we plan to experiment with using existing VL models on natural images of technical objects, such as 2D snapshots of CAD models and online images of products. Since many technical images used by designers are black-and-white and schematic, this experiment will help us evaluate how well current VL models adapt to both natural and schematic technical images. **(3.3)** Our CLIP model is currently evaluated using patents from only three IPC classes (A42B3, A62B18, H02K19), limiting conclusions on model performance to these specific technological domains. For future studies, we plan to expand the dataset to include a broader range of IPC classes. Since the CLIP base model is primarily pre-trained on real-world photographs of everyday objects, it is likely to perform better on patents related to consumer products such as eyeglasses, chairs, and helmets, rather than on highly technical systems like solar batteries, harvesting machines, or industrial furnaces. Conducting further experiments will provide deeper insights into how effectively existing VL models can be fine-tuned for specific technological fields and design challenges. **(3.4)** Our approach relies on examiner citations to automatically build a database of linked patents, eliminating the need for time-consuming manual relevance evaluations. While this method is effective for evaluating the model's ability to retrieve relevant patents, it does not account for the fact that patent citations are backward-looking (i.e., a cited patent must always be older than the citing patent). As a result, our VL model may retrieve subsequent relevant patents. In future studies we plan to adjust the retrieval metrics to ensuring that only prior art is considered for retrieval (Luo et al., 2024).

7. Conclusions

This study presents a review of VL models in the field of ED, examining their application across the ED phases. We identified two key barriers to broader VL adoption in ED: the creation of large-scale multimodal design datasets and the effective representation of ED knowledge. To address these issues, we introduce a scalable and automated process for generating a large-scale multimodal design dataset from patents and an evaluation multimodal dataset using patent citations. Moreover, we fine-tuned CLIP base model and test its performance on patent citations retrieval. Despite the large fine-tuning over a 1.5M of text-image pairs, the poor performances achieved by the models underscores that (1) dataset quality outweighs size in achieving better performance in ED tasks; (2) the need of a deeper understanding of how ED concepts are represented though text and images; and (3) the need for significant research efforts to enable effective representation of ED knowledge. Based on these limitations, we propose potential solutions to guide future research directions. This work is a first attempt to demonstrates that both foundational and fine-tuned VL models exhibit limited readiness for deployment in real-world ED scenarios.

https://doi.org/10.1017/pds.2025.10340 Published online by Cambridge University Press

Acknowledgement

This research has been partly funded by PNRR - M4C2 - Investimento 1.3, Partenariato Esteso PE00000013 - "FAIR - Future Artificial Intelligence Research" - Spoke 1 "Human-centered AI", funded by the European Commission under the NextGeneration EU program and by the DETAILLs Project (DEsign Tools of Artificial Intelligence in Sustainability Living LabS) - European Union. Erasmus + KA2 - Cooperation partnership in higher education (Project Number: 2023-1-IT02-KA220-HED-000158755).

References

BaltruŠaitis,T., Ahuja C., & Morency L. P. (2018). Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence,* 41(2), 423–443. https://doi.org/10.1109/TPAMI.2018.2798607

Consoloni M., Giordano V., & Fantoni G. (2024). Assessing text-image patent datasets with text-based metrics for engineering design applications. *Proceedings of the Design Society,* 4, 1969–1978. https://doi.org/10.1017/pds.2024.199

Edwards K. M., Man B., & Ahmed F. (2024). Sketch2Prototype: rapid conceptual design exploration and prototyping with generative AI. *Proceedings of the Design Society,* 4, 1989–1998. https://doi.org/10.1017/pds.2024.201

Galatolo F. A., Cimino M. G. C. A. & Vaglini G. (2021). Generating Images from Caption and Vice Versa via CLIP-Guided Generative Latent Space Search. *In Proceedings of the International Conference on Image Processing and Vision Engineering, SciTePress, pages* 166–174. https://doi.org/10.5220/0010503701660174

Jin J., Yang M., Hu H., Guo X., Luo J., & Liu Y. (2024). Empowering design innovation using AI-generated content. Journal of Engineering Design, 1–18. https://doi.org/10.1080/09544828.2024.2401751

Kwon E., Huang F., & Goucher-Lambert K. (2022). Enabling multi-modal search for inspirational design stimuli using deep learning. *AI EDAM*, 36, e22. https://doi.org/10.1017/S0890060422000130

Li X., Wang Y., & Sha Z. (2023). Deep learning methods of cross-modal tasks for conceptual design of product shapes: A review. *Journal of Mechanical Design,* 145(4). https://doi.org/10.1115/1.4056436

Lo H. C., Chu J. M., Hsiang J., & Cho C. C. (2024). Large language model informed patent image retrieval. *arXiv preprint arXiv:2404.19360.*

Pan X., Li X., Li Q., Hu Z., & Bao J. (2024). Evolving to multi-modal knowledge graphs for engineering design: state-of-the-art and future challenges. *Journal of Engineering Design*, 1–40. https://doi.org/10.1080/09544828.2023.2301230

Picard C., Edwards K. M., Doris A. C., Man B., Giannone G., Alam M. F., & Ahmed F. (2023). From concept to manufacturing: Evaluating vision-language models for engineering design. *arXiv preprint arXiv:2311.12668*. https://doi.org/10.48550/arXiv.2311.12668

Radford A., Kim J. W., Hallacy C., Ramesh A., Goh G., Agarwal S., ... & Sutskever I. (2021, July). Learning transferable visual models from natural language supervision. *In International conference on machine learning* (pp. 8748–8763). PMLR. https://doi.org/10.48550/arXiv.2103.00020

Siddharth L., Li G., & Luo J. (2022). Enhancing patent retrieval using text and knowledge graph embeddings: a technical note. *Journal of Engineering Design*, 33(8-9), 670–683. https://doi.org/10.1080/09544828.2022.2144714

Song B., Miller S., & Ahmed F. (2023). Attention-enhanced multimodal learning for conceptual design evaluations. *Journal of Mechanical Design,* 145(4), 041410. http://dx.doi.org/10.1115/1.4056669

Song B., Zhou R., & Ahmed F. (2024). Multi-modal machine learning in engineering design: A review and future directions. *Journal of Computing and Information Science in Engineering,* 24(1), 010801. https://doi.org/10.1115/1.4063954

Su H., Song B., & Ahmed F. (2023, August). Multi-modal machine learning for vehicle rating predictions using image, text, and parametric data. *In International Design Engineering Technical Conferences and Computers and Information in Engineering Conference (Vol. 87295, p. V002T02A089). American Society of Mechanical Engineers.* https://doi.org/10.1115/DETC2023-115076

Yuan C., Marion T., & Moghaddam M. (2022). Leveraging end-user data for enhanced design concept evaluation: A multimodal deep regression model. *Journal of Mechanical Design,* 144(2), 021403. https://doi.org/10.1115/1.4052366

https://doi.org/10.1017/pds.2025.10340 Published online by Cambridge University Press