A Appendix

A.1 Baselines

We evaluate several baseline methods for multimodal image retrieval, including both image-only and image-text approaches. The following subsections describe the implementation details.

A.1.1 Implementation Details

For the training of the CLIP model, we used the open source CLIP model training code [Ilharco et al.(2021)]. For the training architecture, we use the vision transformer variant ViT-B-16 as vision encoder and a vanilla transformer model as text encoder. We initialized the weights using a pre-trained model named laion400m_e32, which was pre-trained on 400 million natural image-text pairs. We optimized the model for $33,000$ iterations with early stopping and using AdamW optimizer with a batch size of 256 and a learning rate of $3.54e-6$. The hyperparameters and model architecture were selected using hyperparameter tuning.

We conducted comprehensive hyperparameter optimization across multiple dimensions: (1) **Model architecture**: Vision Transformer variants (ViT-B-16, ViT-B-32) and (2) **Optimization parameters**: Learning rates (10^{-3} to 10^{-6}), weight decay values (0.1, 0.2) for AdamW

Visualizing cross-modal alignment of CLIP After adapting CLIP to the patent domain, we encoded a few patent images and their corresponding figure descriptions using the adapted CLIP model and performed dimensionality reduction of the embeddings using Uniform Manifold Approximation and Projection (UMAP) [McInnes et al.(2018)] to visualize them in a 2D-plane. Figure 1 shows the visualization, where the green circles represent the images, and the blue triangles represent the corresponding figure descriptions in the same projection space. We compare the image-text alignment before (left) and after adaption of CLIP (right).

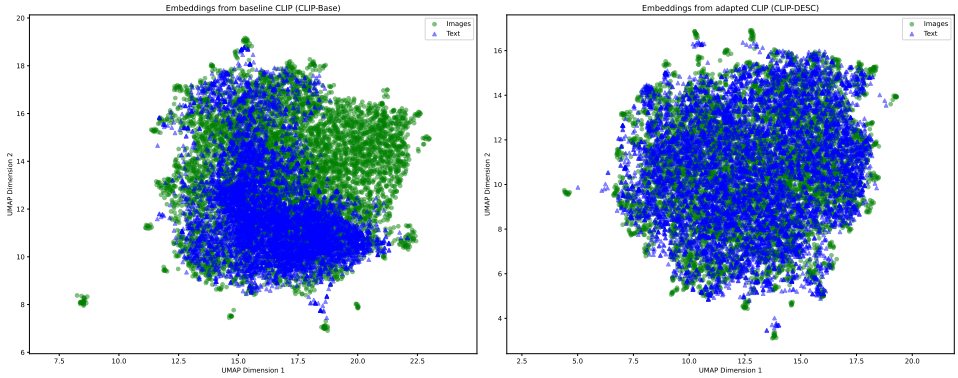


Figure 1: UMAP visualization of CLIP embeddings before and after adaptation to patent domain using figure descriptions for QUERY patents. **Left:** Embeddings from baseline CLIP-*base* model. **Right:** Embeddings from adapted CLIP-*DESC* model. Green circles represent image embeddings, blue triangles represent text embeddings.

A.2 Patent Figure Similarity Annotation Guideline

The objective of this annotation task is to evaluate the relevance between pairs of patent images across three key dimensions. Relevance, in the context of this annotation task, refers to the degree to which the characteristics of one patent image align with the characteristics of another patent image, relative to specified dimensions.

During annotation, each dimension should be considered independently.

A.3 1. Visual Relevance

For visual relevance, we ask the question "How similar are the two shown images visually?" We assess the visual relevance of the two images based on the following points:

- Overall shape and structural match
- Layout and arrangement of components
- Level of detail and complexity
- Ignore leading lines and pointed arrows

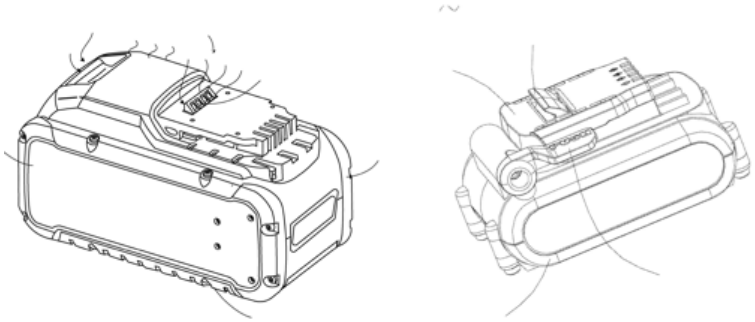
Score	**Condition**
1 - Yes	Nearly identical or very similar visual appearance
0 - No	Completely different visual appearance

Table 1: Score card for annotating visual relevance between shown patent image pairs

Guide: Would these drawings look similar if traced?
Example Here, the following two images are visually similar (1 - Yes) as the overall shape and structure match, although are shown in different perspectives and are different inventions. (first image is a 'battery pack' and second image is 'power brick')

2. Semantic Relevance For semantic relevance, we ask the question "How semantically close are the concepts depicted in the shown two images?" We assess the semantic relevance of the two images based on the following points:

- Functional purpose of the invention
- Core operating principles
- Problem addressed by the invention



(a) Patent image showing *battery pack*

(b) Patent image showing *power brick*

Figure 2: Example of an image pair for visual relevance annotation

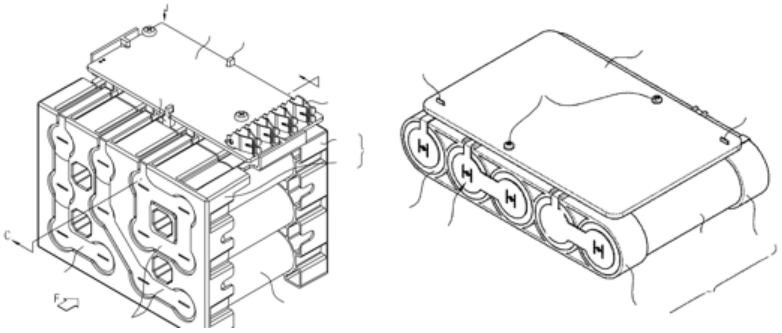
- Application of the invention

Score	**Condition**
1 - Yes	Nearly identical or very similar function and approach
0 - No	Completely different function/purpose

Table 2: Score card for annotating semantic relevance between shown patent image pairs

Guide: Do these inventions solve the same problem?

Example For example, the following two images are showing two different battery modules. Here, there is semantic relevance (1 - Yes) as both inventions have the same function, same working principle and same application, although the images look different. (both images are showing 'battery pack')



(a) Patent image showing *battery pack*

(b) Patent image showing *battery pack*

Figure 3: Example of an image pair for semantic relevance annotation

3. Part-whole Relevance For part-whole relevance, we ask the question "Is one image showing one or more subpart of the same invention depicted in the other image?" We assess the part-whole relevance of the two images based on the following points:

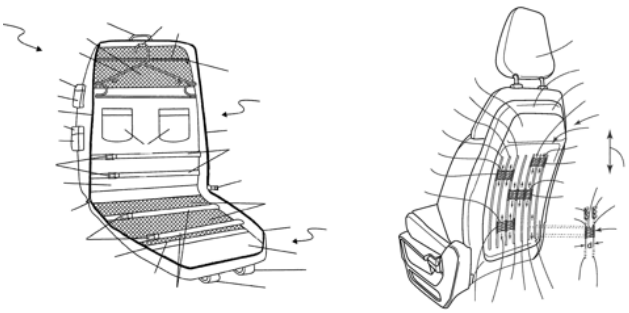
- One image showing key or specific component/s of the invention shown in other image
- Hierarchical relationship exists between the concepts shown in the images

Score	Condition
1 - Yes	One image showing one or more subpart of the other invention
0 - No	No shared components

Table 3: Score card for annotating subpart relevance between shown patent image pairs

Question to Ask Yourself - Is this image a detailed view or component of the other?

Example For example, the following two images are showing two different subparts of the same concept. Here, the images are relevant (1 - Yes) as one is a subpart of the other. (first image is showing 'internal structure of chair' and second image is showing 'back of chair')



(a) Patent image showing *internal structure of a chair*

(b) Patent image showing *back of a chair*

Figure 4: Example of an image pair for semantic relevance annotation

References

[Ilharco et al.(2021)] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. 2021. OpenCLIP. `https://doi.org/10.5281/zenodo.5143773`

[McInnes et al.(2018)] Leland McInnes, John Healy, Nathaniel Saul, and Lukas Grossberger. 2018. UMAP: Uniform Manifold Approximation and Projection. *The Journal of Open Source Software* 3, 29 (2018), 861.