



**HAL**  
open science

# PatFig: Generating Short and Long Captions for Patent Figures

Dana Aubakirova, Kim Gerdes, Lufei Liu

► **To cite this version:**

Dana Aubakirova, Kim Gerdes, Lufei Liu. PatFig: Generating Short and Long Captions for Patent Figures. ICCV workshop: CLVL: 5th Workshop on Closing the Loop Between Vision and Language, Computer Vision Foundation, Oct 2023, Paris, France. pp.2843-2849, 10.1109/ICCVW60793.2023.00305 . hal-04408316

**HAL Id: hal-04408316**

**<https://hal.science/hal-04408316>**

Submitted on 21 Jan 2024

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# PatFig: Generating Short and Long Captions for Patent Figures

Dana Aubakirova<sup>1,2</sup>

Kim Gerdes<sup>1,2</sup>

Lufei Liu<sup>1</sup>

<sup>1</sup>Qatent, Paris, France <sup>2</sup>Université Paris-Saclay, LISN (CNRS), France  
{dana, kim, lufei}@qatent.com

## Abstract

This paper introduces *Qatent PatFig*, a novel large-scale patent figure dataset comprising 30,000+ patent figures from over 11,000 European patent applications. For each figure, this dataset provides short and long captions, reference numerals, their corresponding terms, and the minimal claim set that describes the interactions between the components of the image. To assess the usability of the dataset, we finetune an LVM model on *Qatent PatFig* to generate short and long descriptions, and we investigate the effects of incorporating various text-based cues at the prediction stage of the patent figure captioning process.

## 1. Introduction

Patents are at the economically strategic crossroads of Artificial Intelligence and Intellectual Property, serving as a cornerstone of technical innovation[5]. A pivotal yet largely untapped aspect at the confluence of visual and linguistic analysis is the study of patent figures. These figures are central to the comprehension and elucidation of patent applications, often providing a more efficient medium for conveying complex scientific or technical information than text alone [6, 12]. They comprise technical drawings, block diagrams, flow charts, plots, and grayscale photographs [32]. While prior research has delved into captioning scientific figures, the specific domain of patent figure captioning remains largely unexplored.

We introduce *Qatent PatFig*, a comprehensive patent figure dataset with long and short descriptions, bolstering research in areas like image-to-text, figure-based patent retrieval, figure classification, segmentation, and text-to-image generation. Using *PatFig*, we train image captioning models to aid patent attorneys in improving figure captions. By fine-tuning the Large Vision Language Model *MiniGPT-4* [37] on *PatFig* and adding textual cues during predictions, we strive to boost caption accuracy and patent-specific relevance. The ultimate goal is to connect vision and language, by comprehending visual information designed to facilitate

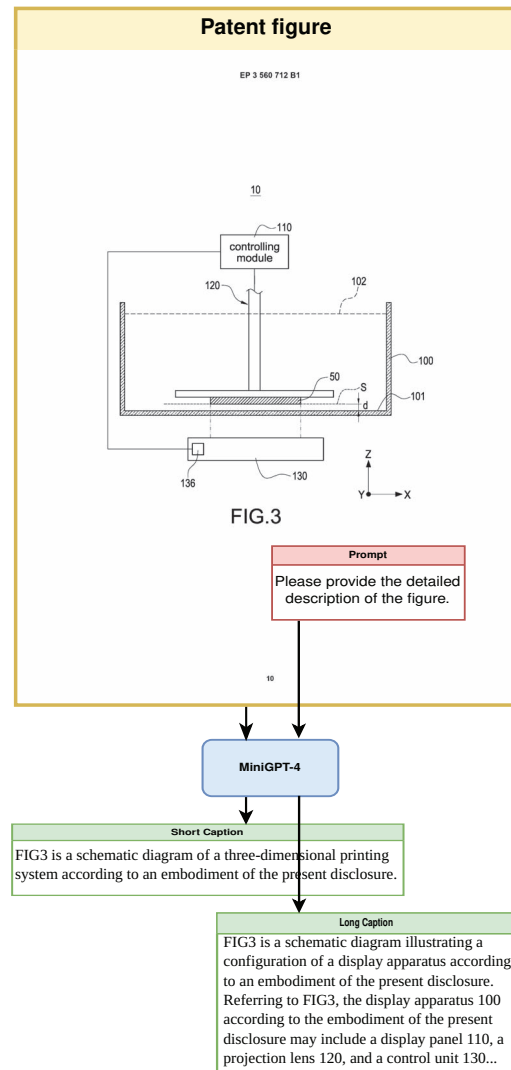


Figure 1. Given an image and a prompt, our figure captioning models generate long and short descriptions. Note that the models are separate for the two types of descriptions.

human cognition of abstract, technical, and scientific concepts. This endeavor is particularly relevant in the context of patent figures, which often encapsulate complex and abstract concepts in a visual format [14].

## 2. Related work

### 2.1. Patent figure datasets

Limited datasets exist for patent figure analysis, primarily targeting image-based patent retrieval. CLEF-IP 2011 [25] provides two such datasets, but with a mere 211 patents and broad image classification across nine categories, it is limited in granularity. The concept dataset [35] has 1000 patent drawings for shoe classification and an additional 2000 mechanical drawings by relevance. Kucer et al.’s DeepPatent [17] offers over 350,000 design patent images<sup>1</sup>. Such patents naturally lack detailed object names, viewpoints, and captions.

We introduce Qatent PatFig, a comprehensive dataset with 30,000 patent figures from 11,000+ patents enriched with long and short descriptions, figure types, reference numerals with terms, and patent claims. While similar datasets like SciCap [13] associate scientific figures with captions, patent figures present unique challenges. They frequently feature reference numerals, term lists, short as well as long descriptions, and the relation between the terms is detailed in the patent claims. However, extracting these descriptions can be daunting due to varied caption structures and the interspersed nature of reference numerals throughout the patent.

### 2.2. Patent figure captioning

Most patent figure research targets figure-based patent querying [17, 31, 25] and classification [14, 36, 19]. For scientific figure captioning, Chen et al. [10, 9, 8] presented FigCAP, using LSTM models with attention. Qian et al.’s FigJAM [26] produces “caption units”, a concept explored earlier with DVQA [15] and FigureQA [16]. SciCap [13] leverages its dataset to train a CNN+LSTM image-captioning model [34].

In this paper, we leverage the recent advancements in Large Vision-Language Models (LVLMs) to address the task of generating short and long captions for patent figures. LLMs, such as LLaMA [29], GPT-3 [22], and Vicuna [11] have demonstrated disruptive progress that can be further extended to large vision-language models [3, 37, 18], thus effectively aligning visual features with the textual space. Yet, their application to the domain of patent figure captioning remains unexplored. We propose to finetune an LVLM to evaluate our dataset and investigate the effectiveness of LVLMs in generating informative and detailed captions for patent figures.

<sup>1</sup>In Europe, “design patents” are termed “Registered Community Design” (RCD). They focus on aesthetic design rather than utility.

## 3. Building the PatFig dataset

In this section, we describe the process of acquiring and pre-processing the data to construct our dataset.

### 3.1. Data acquisition and pre-processing

Qatent’s internal Solr [2] database contains complete textual patent data from the European Patent Office (EPO) including publication number, title, abstract, claim, IPC (patent classification), inventors, patent family, applicants, id, and complete description. Based on this database, we initiated the image acquisition process by retrieving the publication numbers within the time range from January 1, 2020, to December 31, 2020. Subsequently, using Espacenet [1], the EPO’s patent search website, we scraped a total of 62,513 patent images corresponding to 15,645 unique patents based on the patent publication numbers, enabling for accurate linking of the images to their respective textual patent data.

### 3.2. Short and long figure caption extraction

Short descriptions of patent figures usually follow a standard format, separated by new lines, enabling a rule-based extraction method. These descriptions often appear in a section titled “Brief Description of Drawings”. Typical short descriptions start with a figure number and a brief explanation, e.g., “Fig. 1 depicts a bottle pourer per an embodiment.” This uniformity aids in automated extraction of such captions. Our results yielded structured sentences with figure numbers, objects, and viewpoints when available.

Long caption extraction poses more challenges due to the varied structure of patent application descriptions. Addressing this, our method involved text normalization, searching for repeated figure number references, and extracting relevant sections until the start of another paragraph or a different figure mention. We also trimmed overly verbose captions.<sup>2</sup>

### 3.3. Figure-type extraction

We leverage the common structure of short captions and apply rule-based methods to extract key phrases appearing after the “are/is a/an,” “shows,” “illustrates,” and “depicts,” etc. As a result, we retrieved 1506 different classes, which were later reduced to 412 after manual revision and text normalization. We grouped the most frequently appearing categories, ranging from the least abstract (top) to the most abstract (bottom) categories present in the dataset as illustrated in Figure 2 in Annex A.

<sup>2</sup>During caption filtering, statistical analysis determined token count ranges for descriptions. For short captions, the range was 10 to 40 tokens, based on the Interquartile Range (IQR) rule. For long ones, it was 40 to 500 tokens. Descriptions outside these bounds were treated as outliers and excluded.

	PyOCR	EasyOCR	Pytesseract	docTR
Accuracy	14.03	53.22	17.42	72.08
Time (s)	0.49	17.11	0.75	7.02

Table 1. The OCR Performance

### 3.4. Figure-caption matching with OCR

OCR (Optical Character Recognition) is employed to extract reference numerals present in patent images, enabling their matching with corresponding figure descriptions and terms. We evaluated four OCR libraries: PyOCR, EasyOCR, Pytesseract, and docTR [20] to determine the best method for extracting information from patent images. 54 images from eight IPC patent class sections (A, B, C, D, E, F, G, H) were randomly selected and manually annotated with reference numerals and figure numbers. As many patent figures appear sideways, we tested rotating the images at 0, 90, 180, and 270 degrees. Among the libraries, we selected docTR for extracting the numerical information present in the image and matching it with the corresponding description as it achieved the highest accuracy of 72.08 as provided in Table 1.

## 4. Dataset evaluation with LVLm

While traditional approaches such as CNN+RNN models have been prevalent in image captioning tasks, they face limitations in effectively representing features and inhibit training parallelization due to the recurrent nature of RNNs [33]. Thus, given the domain-specific and technical nature of patent figures, relying solely on CNN+RNNs may not capture the intricate semantic relationships and contextual nuances required for generating accurate captions. Following the recent trends in utilizing autoregressive language models for vision-language tasks, capitalizing on cross-modal transfer, we chose a large vision-language image-captioning model, MiniGPT-4, as our baseline model. We created three variations of the baseline models, namely Vision-only and two versions of Vision+Text. We evaluate the generated short and long captions using common evaluation metrics: BLEU [24], ROUGE[21], METEOR [4], and CIDEr [30] scores.

### 4.1. Experimental setup

We finetuned the MiniGPT-4 model separately on short and long captions. For both models, we trained for 10 epochs, setting the maximum text length to 50 for short captions and 500 for long captions. The fine-tuning for short captions was carried out on a 1xRTX A6000 GPU, while the long captions were fine-tuned on a 1xA100 80GB GPU. MiniGPT-4 is composed of a vision encoder with a pre-trained Vision Transformer (ViT) and Q-Former, a single linear projection layer, and an advanced Vicuna large lan-

The number of images with captions	Raw dataset ( $\geq 1$ figures per image)	Revised dataset (1 figure per image)	
		Train	Test
	30714	17877	2417

Table 2. The PatFig dataset statistics

	Input			Metric				
	Image	Title	Terms	BLEU2	BLEU4	ROUGE	METEOR	CIDEr
Short captions				with reference numerals				
	+	-	-	0.5322	0.3851	0.3677	0.3336	0.3669
	+	+	-	0.5537	0.4206	0.4007	0.3753	0.7947
	+	+	+	0.2202	0.1553	0.2358	0.2232	0.1903
				without reference numerals				
	+	-	-	0.5359	0.3922	0.4071	0.3640	0.3413
Long Captions	+	+	-	0.5573	0.4276	0.4390	0.4056	0.7939
	+	+	+	0.2228	0.1577	0.2577	0.2338	0.1105
				with reference numerals				
	+	-	-	0.3281	0.1936	0.1949	0.1410	0.0114
	+	+	-	0.3437	0.2162	0.2212	0.1595	0.0366
	+	+	+	0.3255	0.2166	0.2308	0.1740	0.0582
			without reference numerals					
+	-	-	0.3313	0.1974	0.2317	0.1534	0.0142	
+	+	-	0.3478	0.2210	0.2598	0.1726	0.0418	
+	+	+	0.3267	0.2177	0.2552	0.1755	0.0587	

Table 3. Results for various captioning configurations, assessed both with and without reference numerals. Experiments using terms were done on test subsets with retrievable non-empty term lists.

guage model. Fine-tuning involved training only the linear projection layer to align the visual features with the Vicuna model.

The authors of MiniGPT provide 2 stages of pretraining. In the first pretraining stage, MiniGPT-4 acquires vision-language knowledge from a combined dataset of Conceptual Caption [28, 7], SBU [23], and LAION [27]. However, it exhibited issues with the coherence of the generated texts. To address this, they provide the second-stage finetuning on a smaller highly curated dataset to refine the generated descriptions. We finetune on the second stage. The qualitative results are provided in 4 in B. The fine-tuning process utilized predefined prompts such as: “Please provide the detailed description of the figure” or “Describe the contents of the image in detail.” The finetuned MiniGPT-4 model on aligned patent figures with short and long captions exhibited the ability to produce more specific and contextually appropriate descriptions.<sup>3</sup>

### 4.2. Vision only

**Task:** Given an image and general prompt generate the description

**Prompt 1:** *Please provide the detailed description of the figure.*

The experiment consists of two main components: the

<sup>3</sup>Notably, we observed that the finetuning process was not efficient, requiring 4500 training steps with a batch size of 4, which is the maximum batch size suitable for the dataset with long captions (max. 500 tokens) that fits within a single 80GB A100 GPU. The multiGPU options for the finetuning stage have not been released yet.

input image and a simple prompt. The input image is fed into the model for processing, and the model generates a caption of the image based solely on its visual content. The goal of this experiment is to evaluate the model’s ability to understand and describe images without any additional text-based context.

### 4.3. Vision+Text

**Task:** Given an image and a prompt including the patent title generate the description.

**Title:** “Activation of energy devices”

**Terms:** 137602: “sensor”, 137604: “wired connection”, 137650: “surgical site opening”, 120: “patient side cart”, 137606: “surgical instrument”, 137600: “retractor”.

**Prompt 2:** *Please provide the detailed description of the figure associated with title.*

**Prompt 3:** *Please provide the detailed description of the figure associated with title. The image contains the following reference numerals and terms.*

This experiment aims to evaluate the impact of added text-based cues on generating contextually accurate captions. The model uses the image as visual input and incorporates the title and terms as additional text-based context for more detailed and relevant captioning. The terms are retrieved from the patent application’s complete description, and a subset of 500 samples from the test data is selected to assess the terms’ effect.

## 5. Discussion

Long descriptions usually mention the different parts that the numbers in the figure refer to; short captions generally do not. As expected, the long caption generation generally benefits from adding the terms to the input, in particular for the CIDEr score, conceived as a caption metric (except for the BLEU2 score). Interestingly, the short caption generation seems to be disturbed by the term list in the input. This is in line with results by [13] on scientific image captioning. Their BLEU scores are similarly low, e.g. at 0.0231 for vision-only models, and decrease even further when adding textual cues. This might be explained by the fact that the model may overlook crucial visual features, resulting in less accurate captions. Additionally, conflicts between textual and visual cues may also confuse the model.

It is well-known that neural vision models do not implicitly learn OCR, and with the goal of generating good captions, it is not actually necessary that our model generates the reference numerals as we could simply add the extracted reference numerals to the matching terms in the generated caption. So the model’s performance is generally better when evaluating without taking the numerals into account.

The significant difference between the gold corpus and the generated captions can be attributed to limitations of

MiniGPT-4: 1) The use of a frozen Q-former in the visual encoder may result in the loss of key features like visual-spatial grounding. 2) Training only a single projection layer may limit the model’s ability to learn comprehensive visual-text alignment effectively.

## 6. Conclusion and future work

This paper introduces the first extensive dataset for gauging the efficiency of Large Vision Language Models (LVLM) on patent figures. Distinguished by reference numerals, a formal template-like caption style, and a wealth of text data linked to each figure, this dataset provides a unique challenge, differentiating it from conventional captioning tasks. Moreover, PatFig encapsulates wider types of patent images compared to the existing datasets, spanning technical drawings, block diagrams, flow charts, plots, and grayscale photographs. It offers multiple data points that can be harnessed for image captioning and potentially for other tasks such as image search and image generation, as well as addressing a broader scope of patent figure analysis tasks.

We delved into a key function of LVLM, exploring the dynamic interplay between language and image during the generation of two distinct caption types: short and long. These variants necessitate different input information, and our findings affirm that our LVLM models can effectively assimilate this information, thereby enhancing the results. Yet, another variations of our experiments should be studied, such as variations of training data size, prompt improvements, with the inclusion of textual cues during the finetuning stage, training directly on texts from which reference numerals have been removed, and, more difficult, removing the numbers from the images. Additionally, the identified patent figure types can be used to categorize the results based on each figure type.

An intriguing aspect warranting further study is identifying the threshold where the image itself becomes redundant in the text generation process. In other words, discerning when a text-only large language model can accurately predict the figure’s content without directly analyzing it.

Future research will expand to investigate the generation of figures from patent text. This could not only streamline the work of patent attorneys significantly but also shed light on the necessary information for drawing a figure and how this data is amalgamated to create a figure.

## 7. Acknowledgments

The authors would like to thank the rest of the Qatent team, including all researchers, engineers, developers, and law experts, for their insights and collaboration throughout the project.

## References

- [1] European Patent Office – Espacenet Patent Search. <https://worldwide.espacenet.com/patent/> Accessed on August 19, 2023. 2
- [2] Apache solr. <https://solr.apache.org>. Accessed on August 19, 2023. 2
- [3] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022. 2
- [4] Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72, 2005. 3
- [5] Richard S Campbell. Patent trends as a technological forecasting tool. *World patent information*, 5(3):137–143, 1983. 1
- [6] Russell N Carney and Joel R Levin. Pictorial illustrations still improve students’ learning from text. *Educational psychology review*, 14:5–26, 2002. 1
- [7] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3558–3568, 2021. 3
- [8] Charles Chen, Ruiyi Zhang, Sungchul Kim, Scott Cohen, Tong Yu, Ryan Rossi, and Razvan Bunescu. Neural caption generation over figures. In *Adjunct Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2019 ACM International Symposium on Wearable Computers*, pages 482–485, 2019. 2
- [9] Charles Chen, Ruiyi Zhang, Eunyee Koh, Sungchul Kim, Scott Cohen, and Ryan Rossi. Figure captioning with relation maps for reasoning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1537–1545, 2020. 2
- [10] Charles Chen, Ruiyi Zhang, Eunyee Koh, Sungchul Kim, Scott Cohen, Tong Yu, Ryan Rossi, and Razvan Bunescu. Figure captioning with reasoning and sequence-level training. *arXiv preprint arXiv:1906.02850*, 2019. 2
- [11] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%\* chatgpt quality. See <https://vicuna.lmsys.org> (accessed 14 April 2023), 2023. 2
- [12] Debasis Ganguly, Johannes Leveling, and Gareth JF Jones. United we fall, divided we stand: A study of query segmentation and prf for patent prior art search. In *Proceedings of the 4th workshop on Patent information retrieval*, pages 13–18, 2011. 1
- [13] Ting-Yao Hsu, C Lee Giles, and Ting-Hao’Kenneth’ Huang. Scicap: Generating captions for scientific figures. *arXiv preprint arXiv:2110.11624*, 2021. 2, 4
- [14] Shawn M Jones and Diane Oyen. Abstract images have different levels of retrievability per reverse image search engine. In *European Conference on Computer Vision*, pages 203–222. Springer, 2022. 1, 2
- [15] Kushal Kafle, Brian Price, Scott Cohen, and Christopher Kanan. Dvqa: Understanding data visualizations via question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5648–5656, 2018. 2
- [16] Samira Ebrahimi Kahou, Vincent Michalski, Adam Atkinson, Ákos Kádár, Adam Trischler, and Yoshua Bengio. Figureqa: An annotated figure dataset for visual reasoning. *arXiv preprint arXiv:1710.07300*, 2017. 2
- [17] Michal Kucer, Diane Oyen, Juan Castorena, and Jian Wu. Deepatent: Large scale patent drawing recognition and retrieval. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2309–2318, 2022. 2
- [18] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023. 2
- [19] Shaobo Li, Jie Hu, Yuxin Cui, and Jianjun Hu. Deepatent: patent classification with convolutional neural networks and word embedding. *Scientometrics*, 117:721–744, 2018. 2
- [20] Haofu Liao, Aruni RoyChowdhury, Weijian Li, Ankan Bansal, Yuting Zhang, Zhuowen Tu, Ravi Kumar Satzoda, R Manmatha, and Vijay Mahadevan. Doctr: Document transformer for structured information extraction in documents. *arXiv preprint arXiv:2307.07929*, 2023. 3
- [21] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004. 3
- [22] Andrea Madotto, Zihan Liu, Zhaojiang Lin, and Pascale Fung. Language models as few-shot learner for task-oriented dialogue systems. *arXiv preprint arXiv:2008.06239*, 2020. 2
- [23] Vicente Ordonez, Girish Kulkarni, and Tamara Berg. Im2text: Describing images using 1 million captioned photographs. *Advances in neural information processing systems*, 24, 2011. 3
- [24] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002. 3
- [25] Florina Piroi, Mihai Lupu, Allan Hanbury, and Veronika Zenz. Clef-ip 2011: Retrieval in the intellectual property domain. In *CLEF (notebook papers/labs/workshop)*, 2011. 2
- [26] Xin Qian, Eunyee Koh, Fan Du, Sungchul Kim, Joel Chan, Ryan A Rossi, Sana Malik, and Tak Yeon Lee. Generating accurate caption units for figure captioning. In *Proceedings of the Web Conference 2021*, pages 2792–2804, 2021. 2
- [27] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo

Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021. 3

- [28] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, 2018. 3
- [29] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 2
- [30] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015. 3
- [31] Stefanos Vrochidis, Anastasia Mourtzidou, and Ioannis Kompatsiaris. Concept-based patent image retrieval. *World Patent Information*, 34(4):292–303, 2012. 2
- [32] Xin Wei, Jian Wu, Kehinde Ajayi, and Diane Oyen. Visual descriptor extraction from patent figure captions: A case study of data efficiency between bilstm and transformer. In *Proceedings of the 22nd ACM/IEEE Joint Conference on Digital Libraries*, pages 1–5, 2022. 1
- [33] Congcong Wen, Yuan Hu, Xiang Li, Zhenghang Yuan, and Xiao Xiang Zhu. Vision-language models in remote sensing: Current progress and future trends. *arXiv preprint arXiv:2305.05726*, 2023. 3
- [34] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057. PMLR, 2015. 2
- [35] Longhui Zhang, Lei Li, and Tao Li. Patent mining: a survey. *ACM Sigkdd Explorations Newsletter*, 16(2):1–19, 2015. 2
- [36] Liyuan Zhang, Wei Liu, Yufei Chen, and Xiaodong Yue. Reliable multi-view deep patent classification. *Mathematics*, 10(23):4545, 2022. 2
- [37] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023. 1, 2

## A. Figure type extraction

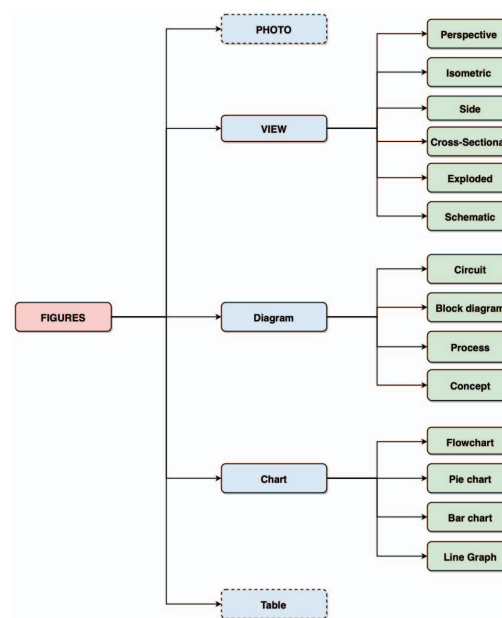


Figure 2. The taxonomy of patent figure types.

## B. Model predictions

Patent number	EP3554181B1
Image id	EP_3554181_B1_4.png
Title	FRONT-HAUL TRANSPORT NETWORK, DATA TRANSMISSION METHOD, APPARATUS AND COMPUTER STORAGE MEDIUM
Terms	1201: "selecting unit", 1202: "first sending unit"
Minimal claim set (Claims covering the maximum number of relevant terms)	15. A device for data transmission, characterized in that it is configured to control data transmission of a front-haul transport network, FTN, wherein the FTN comprises: an access-type front-haul transport node, FTN-ACC, connected to a remote radio unit, RRU; and an aggregation-type front-haul transport node, FTN-AGG, connected to the FTN-ACC and a base band unit, BBU, pool respectively, wherein the FTN-ACC is a node separate from the RRU and the FTN-AGG is a node separate from the BBU pool; wherein the device for data transmission is the FTN-ACC or the FTN-AGG, and comprises: a selecting unit (1201), configured to select, according to a delay requirement of data, a transmission channel established between the FTN-ACC and the FTN-AGG; and a first sending unit (1202), configured to send the data by using the transmission channel.
Short Caption	Long Caption
Groundtruth	
FIG12 illustrates a schematic structural diagram of a device for data transmission of an FTN according to an example of the present disclosure	A device for data transmission is further provided in this example, which is configured to control data transmission of an FTN. The FTN includes an FTN-ACC connected to the RRU, and an FTN-AGG connected to the FTN-ACC and a BBU pool respectively. As illustrated in FIG12, the device for data transmission includes a selecting unit 1201 and a first sending unit 1202.
Predicted: Image	
FIG12 is a schematic structural diagram of a terminal according to an embodiment of the present invention	In an embodiment of the present invention, the network device may be a communication device. The communication device may be a base station, a relay station, or a user equipment. FIG12 is a schematic structural diagram of a communication device according to an embodiment of the present invention. As shown in FIG12, the communication device may include a processor 1201, a memory 1202, and a transceiver 1203. The processor 1201 may be configured to implement the foregoing methods of the embodiments of the present invention. The memory 1202 may be configured to store the foregoing methods of the embodiments of the present invention. The transceiver 1203 may be configured to send and receive signals.
Predicted: Image + Title	
FIG13 is a schematic structural diagram of a base station according to an embodiment of the present invention	FIG1 is a schematic structural diagram of a front-haul transport network according to an embodiment of the present invention. As shown in FIG1, the front-haul transport network may include a first base station and a second base station. The first base station and the second base station may be connected to each other through a first optical fiber and a second optical fiber, respectively. The first optical fiber and the second optical fiber may be connected to each other through a third optical fiber. The first base station and the second base station may be connected to each other through the third optical fiber.
Predicted: Image + Title + Terms	
This is a diagram of a front-haul transport network and a data transmission method, apparatus, and computer storage medium thereof according to an embodiment of the present invention	FIG12 is a schematic structural diagram of an apparatus for transmitting data according to an embodiment of the present invention. As shown in FIG12, the apparatus 1200 for transmitting data may include a selecting unit 1201 and a first sending unit 1202.

Table 4. Test set data sample with associated title, patent number, a minimal claim set, ground truth, and predicted short and long captions.