PatentVision: A multimodal method for drafting patent applications

Ruo Yang
Samsung Semiconductor, Inc.
San Jose, CA
r.yang@partner.samsung.com

Sai Krishna Reddy Mudhiganti
Samsung Semiconductor, Inc.
San Jose, CA
s.mudhiganti@samsung.com

Manali Sharma
Samsung Semiconductor, Inc.
San Jose, CA
manali.s@samsung.com

Abstract

Patent drafting is complex due to its need for detailed technical descriptions, legal compliance, and visual elements. Although Large Vision-Language Models (LVLMs) show promise across various tasks, their application in automating patent writing remains underexplored. In this paper, we present PatentVision, a multimodal framework that integrates textual and visual inputs—such as patent claims and drawings—to generate complete patent specifications. Built on advanced LVLMs, PatentVision enhances accuracy by combining fine-tuned vision-language models with domain-specific training tailored to patents. Experiments reveal it surpasses text-only methods, producing outputs with greater fidelity and alignment with human-written standards. Its incorporation of visual data allows it to better represent intricate design features and functional connections, leading to richer and more precise results. This study underscores the value of multimodal techniques in patent automation, providing a scalable tool to reduce manual workloads and improve consistency. PatentVision not only advances patent drafting but also lays groundwork for broader use of LVLMs in specialized areas, potentially transforming intellectual property management and innovation processes.

1 Introduction

Drafting a comprehensive patent specification involves transforming intricate technical concepts, embodied in both written claims and accompanying illustrations, into precise and coherent legal documentation. Traditional methods predominantly focus on textual analysis, leveraging natural language processing techniques to interpret and generate patent specifications. However, these approaches often overlook the critical role of visual elements—patent drawings—which serve as indispensable carriers of design intent and functional details. As a result, existing systems struggle to fully capture the nuanced interplay between textual and visual components, leading to limitations in accurately reflecting inventors' intentions and meeting professional drafting standards. In recent years, advances in Large Vision-Language Models (LVLMs) have demonstrated significant potential in bridging the gap between linguistic and visual domains. By integrating multimodal data streams, LVLMs enable a deeper comprehension of contextually rich scenarios, offering new avenues for enhancing automated processes across diverse applications. This study investigates the application of state-of-the-art LVLMs, including models such as Gemma (Team et al., 2025), LLAVA (Liu et al., 2024), and LLaMA (Grattafiori et al., 2024), to address the challenges inherent in patent specification drafting. Specifically, we examine how these models can effectively combine patent claims and corresponding drawings to produce high-quality patent specifications. Through rigorous experimentation, our findings reveal that incorporating visual inputs significantly elevates the accuracy and coherence of generated texts, closely mirroring established human drafting practices.

The proposed framework employs a dual-input architecture, where textual inputs consist of patent claims and descriptive annotations, while visual inputs encompass detailed patent diagrams. By fusing these modalities, the system achieves a holistic interpretation of the invention, enabling it to generate specifications that are not only technically accurate but also aligned with legal requirements. These insights underscore the transformative potential of multimodal approaches in automating patent drafting, paving the way for more efficient and reliable intellectual property management.

2 Related Work

Most prior work on patent text generation has focused on specific sections rather than full specifica-

Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 5: Industry Track), pages 397–405
March 25-27, 2026

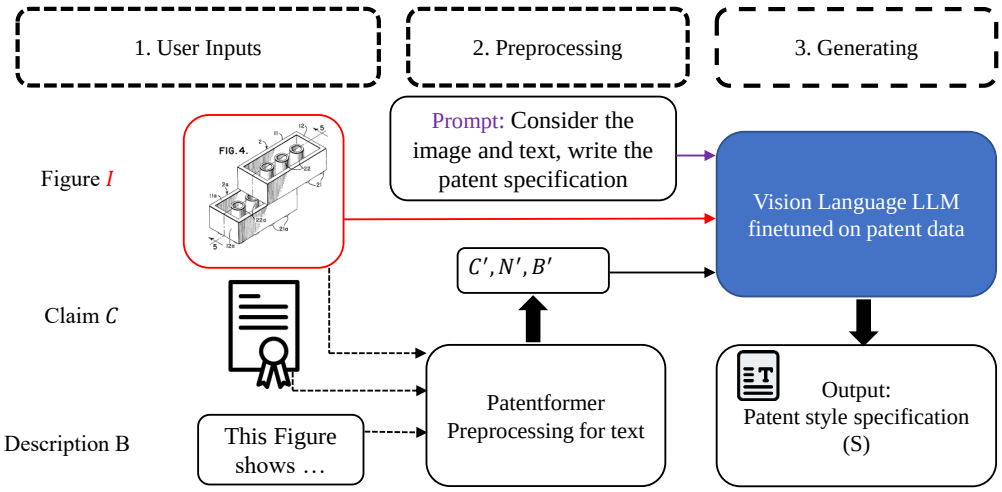

Figure 1: PatentVision is a framework that generates high-quality patent specifications using multimodal inputs like images, patent claims, and optional figure descriptions. Specifically, PatentVision integrates three inputs: the image, enriched textual content derived from PatentFormer's text processing pipeline (Wang et al., 2024), and an instruction prompt tailored for the base vision-language model. The vision-language model is fine-tuned on domain-specific patent data to learn and replicate the formal writing style typical of patent specifications, thereby assisting patent authors in drafting coherent and contextually appropriate descriptions.

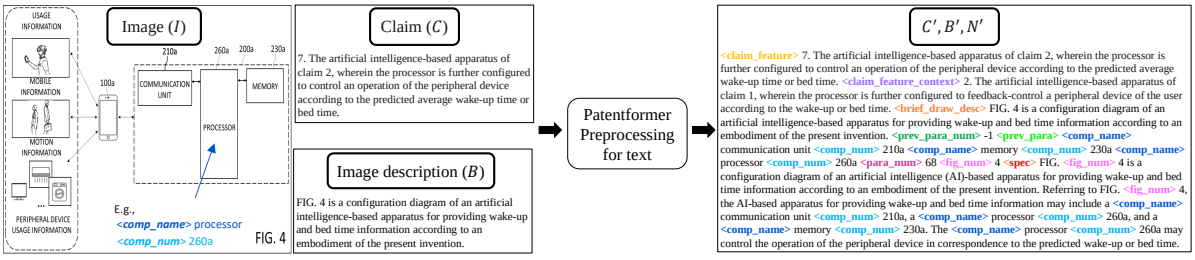

Figure 2: PatentFormer (Wang et al., 2024) performs text processing by taking as input the image I, the claim C, and the image description B. It outputs an enriched textual representation containing structured tokens such as $<comp\_name>$, which are subsequently encoded using the tokenizer of the language model. These enriched tokens provide explicit semantic anchors that facilitate more accurate and context-aware specification generation.

tions. For example, Lee and Hsiang (2020a) fine-tuned GPT-2 to generate claims; Lee (2020c) added a BERT-based module for personalized claim generation; Lee and Hsiang (2020b) introduced a span-based framework for evaluating claim generation; and Jiang et al. (2024) generated claims from detailed descriptions. Lee (2020a) used structural metadata to control generation via text-to-text mappings, while Lee (2020b) applied semantic search for control. Lee (2023) pre-trained GPT-J on patent corpora for autocompletion and introduced the Autocomplete Effectiveness (AE) ratio, which Jieh-Sheng (2022) extended using bidirectional pre-training of GPT-J-6B. Christofidellis et al. (2022) proposed Patent Generative Transformer (PGT), a GPT-2-based model for part-specific generation. Other work focused on summarizing patents to produce titles (Souza et al., 2021), abstracts (Guoliang et al., 2023; Zhu et al., 2023), prior art (Lee and Hsiang, 2020c), or figure captions (Aubakirova et al., 2023). Separately, research on modeling long documents in legal and medical domains has explored hierarchical transformers and efficient attention mechanisms, such as Longformer (Beltagy et al., 2020), Linformer (Wang et al., 2020), Big Bird (Zaheer et al., 2020), and Hi-Transformer (Wu et al., 2021). BioGPT (Luo et al., 2022) fine-tuned GPT-2 for biomedical tasks. Li et al. (2024) provide a survey on pretrained language models for long-form generation.

In contrast to prior work focused on generating short sections or summaries, our approach builds on PatentFormer (Wang et al., 2024) and is, to our knowledge, the first work to generate full patent

specifications directly from claims and drawings.

3 Methodology

Formally, let $\mathcal{P}$ represent a patent document containing a sequence of l claims, $\mathcal{C} = \{c_1, c_2, ..., c_l\}$, a sequence of m specification paragraphs, $\mathcal{S} = \{s_1, s_2, ..., s_m\}$, a set of t drawing images, $\mathcal{I} = \{i_1, i_2, ..., i_t\}$, and a set of t brief descriptions of the drawings, $\mathcal{B} = \{b_1, b_2, ..., b_t\}$, corresponding to each image in $\mathcal{I}$. For $\forall i_z \in I$, let n_z represent a set of k pairs of component names and their respective component numbers that appear in the drawing; $n_z = \{<i_{z_1}^{name}, i_{z_1}^{num}>, <i_{z_2}^{name}, i_{z_2}^{num}>, ..., <i_{z_k}^{name}, i_{z_k}^{num}>\}$, where $i_{z_j}^{name}$ is the name of j^{th} component and $i_{z_j}^{num}$ is the number of j^{th} component in image i_z; $\mathcal{N} = \{n_1, n_2, ..., n_t\}$ corresponding to all images in $\mathcal{I}$. Each image is preprocessed by first rotating it to the correct orientation and then rescaling it such that the maximum of its height or width is 4096 pixels. In a later section, we analyze the impact of image resolution on model performance and demonstrate that higher-resolution images lead to improved specification generation compared to lower-resolution settings.

3.1 Claim+Diagram-to-Specification

Instead of generating specifications from text input only in (Wang et al., 2024), e.g., *text-to-text*, we introduce a multimodal task (*image-text-to-text*) called claim+diagram-to-specification, $\mathcal{T} \rightarrow \mathcal{S}$. Its goal is to generate output specification, $\mathcal{S}$, by using $\mathcal{C}$, $\mathcal{B}$, $\mathcal{I}$, and $\mathcal{N}$ as inputs, where the output specification must support all the input claim features, $\mathcal{C}$, correctly describe the drawings by using drawing descriptions, $\mathcal{B}$, the corresponding image associated with the claim, $\mathcal{I}$, and pairs of components, $\mathcal{N}$, associated with each drawing.

We construct training samples containing the input and output pairs, $<\mathcal{T}, \mathcal{S}>$, where $\mathcal{T}=<\mathcal{C}, \mathcal{B}, \mathcal{I}, \mathcal{N}>$. Rather than learning from all the input text at once to produce the entire specification, we introduce an auxiliary task of mapping each claim feature to a paragraph in the specification and use only one drawing[1] associated with a paragraph. We first match b_z to s_y by checking for common figure numbers. Then, we match s_y to c_x by using the average of cosine similarity and BLEU scores between s_y and c_x. Each $s_y \in \mathcal{S}$ may describe a figure or not. We only keep paragraphs that describe at least one figure in the patent by checking the presence of the words 'FIG.', 'Fig.', and 'Figure', as well as occurrences of any component names and numbers in each paragraph. To simulate the extraction of component names and numbers from a drawing image i_z in the training data, we extract n_z from each s_y, as described in (Wang et al., 2024).[2] Finally, we construct the quadruplets of samples, $<c_x, b_z, i_z, n_z, s_y>$, where $<c_x, b_z, i_z, n_z>$ is the input to produce the corresponding output specification, s_y. We customize the tokenizer and insert special tags into the input and output tokens to help the model understand different contexts.

3.2 PatentVision

Now we introduce our multimodal model, PatentVision, that embeds rich context into the training data for generating specifications and uses patent images directly. Similar to (Wang et al., 2024), first, for each claim feature extracted from an independent claim, we provide as context the remaining claims features of that claim, and for each claim feature extracted from a dependent claim, we provide as context any remaining claim features as well as its parent claim as context. Second, for each figure number, component name, and component number, we embed special tags in both the input and output specifications to mark their presence in the training data. Third, we also provide context by referencing the previous paragraph number and the current paragraph number to help the model understand the context and generate a coherent specification. As an real example shown in Figure 2, we represent the enriched versions of $\mathcal{C}$, $\mathcal{N}$, and $\mathcal{S}$ as $\mathcal{C}'$, $\mathcal{N}'$, $\mathcal{S}'$, and $\mathcal{B}'$=$\mathcal{B}$. As shown by (Wang et al., 2024), embedding rich context into the training data yields significant improvements in the model's performance.

Instead of relying solely on textual input, PatentVision integrates multimodal vision-language models to improve specification quality by incorporating both visual and textual information. PatentVision extends the capabilities of PatentFormer (Wang et al., 2024) in two key as-

[1] Note that some paragraphs may describe more than one drawing. In this work, we assume that each paragraph describes only one drawing, and remove the lines from paragraph that refer to other figures.

[2] USPTO provides patent drawings in .TIFF or .PDF formats, so the extraction of component names and numbers from images is not accurate; hence, we simulated the extraction of component names and numbers from specification, instead. In practice, the drawing files are usually provided in Visio or powerpoint formats, from which extracting the component names and numberings is straightforward.

pects. First, it interprets and utilizes visual content from figures associated with patent claims to enhance the generation of specifications by jointly modeling visual and textual modalities. Second, unlike PatentFormer, which generates outputs solely based on enriched text inputs, PatentVision is designed as an interactive agent capable of engaging in dialogue with the human users. Specifically, it accepts human instructions, enriched textual descriptions, and visual inputs as part of the specification generation process. As a result, the generated specification can vary according to the provided human instructions, enabling greater flexibility.

4 Experimental Setup

In this section, we provide details of the experimental settings, including the dataset, models, baselines, evaluation metrics, and hardware specifics used for training and evaluation of PatentVision.

Dataset. We construct the first dataset for generating specifications from the claims and associated drawings. We worked with four patent experts and focused on generating patents for a specific CPC code, 'G06F'[3] , which includes patents from a diverse range of topics related to electronic digital data processing. The dataset contains a total of 230K image-text-to-text samples. Due to the high computational cost of inference during evaluation, we randomly sample 1K instances as the test set, while the remaining samples are used for training.

Models. To train PatentVision, we evaluate three large vision-language models (LVLMs) as its core components: Gemma 3-12B (Team et al., 2025), LLAVA 1.6-13B (Liu et al., 2024), and LLaMA 3.2-11B (Grattafiori et al., 2024). Each model is fine-tuned on the `Patent-2015-2023-G06F` dataset. Based on empirical performance, the best-performing model (Gemma 3) is selected for deployment within the PatentVision framework.

Baselines. To the best of our knowledge, PatentFormer (T5-11B (Raffel et al., 2020)) is the first work that addresses the task of generating specifications from both patent claims and corresponding drawings. As there is no prior baseline in the literature for direct comparison, we evaluate the performance of PatentVision against PatentFormer, the most closely related approach. For a fair comparison, we adopt the same post-processing strategy as described in (Wang et al., 2024), which ranks generated paragraphs based on alignment with input claims, component names, component numbers, and the correct figure number. The top-ranked paragraph is then selected as the final output.

Evaluation Metrics. To compare the models under various settings, we report the performance of PatentVision using ten popular metrics for natural language generation from the literature, including Bertscore (Zhang* et al., 2020), BLEU score (Papineni et al., 2002; Lin and Och, 2004), ROUGE scores (R-1, R-2, R-L, and R-Lsum) (Lin, 2004), WER (Woodard and Nelson, 1982), Chrf (Popović, 2017, 2015), METEOR (Banerjee and Lavie, 2005), and NIST (Doddington, 2002).

Training. We utilized NVIDIA A100 GPUs (80 GB per GPU) for model training. Each model was trained for 1 epoch. Rather than fine-tuning the VL models directly, which requires a significant amount of GPU resources, we choose to fine-tune the models using LoRA (Hu et al., 2022) instead.

5 Experimental Results

In this section, we begin by comparing the performance of the multimodal PatentVision with the text-only PatentFormer to assess the benefits of incorporating visual understanding into the patent specification generation task. Next, we evaluate large vision-language models (LVLMs) on a dataset without image descriptions to demonstrate that PatentVision produces higher-quality outputs than PatentFormer, even when requiring less human input. We then compare fine-tuned LVLMs with their pretrained counterparts to quantify the quality improvements achieved through fine-tuning on our patent dataset. Finally, we examine the performance of LVLM across different training epochs and image resolutions to analyze the sensitivity of LVLMs to key hyperparameters. We additionally include evaluation tables in the appendix.

5.1 PatentVision vs. PatentFormer

To evaluate the benefits of incorporating visual understanding into the patent specification generation task, we first compare the performance of PatentVision, instantiated with different multimodal models, against the text-only PatentFormer. Specifically, we fine-tune PatentVision using Gemma 3, LLAVA 1.6, and LLaMA 3.2 as base models, each with varying LoRA ranks, on the Image-Text-to-Text patent data pairs. In parallel, PatentFormer is fine-tuned on the same dataset, but without access to image

[3] https://www.uspto.gov/web/patents/classification/cpc/html/defG06N.html#G06F

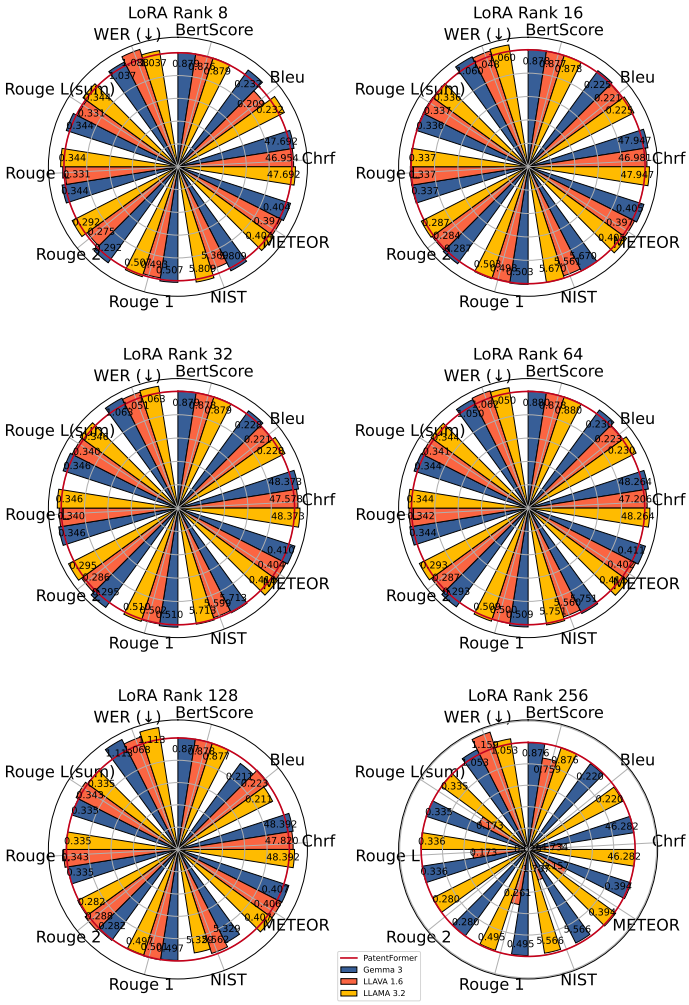

Figure 3: Comparison between PatentVision with different base LVLMs and LoRA ranks and PatentFormer.

inputs. All models are trained for a single epoch to ensure consistent conditions. Figure 3 presents the results comparing PatentVision (with different base LVLMs) to PatentFormer. As shown, PatentVision consistently outperforms PatentFormer across all evaluation metrics.

5.2 Finetuned vs. pretrained VL models

Next, we evaluate the capability of original pretrained vision-language (VL) models on the patent specification generation task without any fine-tuning. This experiment allows us to assess the extent to which fine-tuning on our patent dataset improves model performance for domain-specific writing. Figure 4 presents the results of both pretrained and fine-tuned VL models, where fine-tuning is performed with a LoRA rank of 64. The results clearly demonstrate that fine-tuned models substantially outperform their pretrained counterparts across all evaluation metrics, particularly in generating specifications consistent with legal and technical writing conventions.

5.3 Removing image descriptions

Based on previous results, Gemma 3 outperforms both LLAVA 1.6 and LLaMA 3.2 on the patent specification generation task. Therefore, we focus subsequent analyses on PatentVision instantiated with Gemma 3 as the base model. We evaluate both PatentFormer and PatentVision (with Gemma 3) on the test dataset without any image descriptions , B. This setting allows us to assess whether PatentVision can learn to interpret visual content directly from raw images. As shown in Figure 6, as expected, removing the image description results in a slight performance degradation due to the reduced input information. However, PatentVision still significantly outperforms PatentFormer in the absence of image descriptions. Notably, PatentVision without image descriptions achieves better

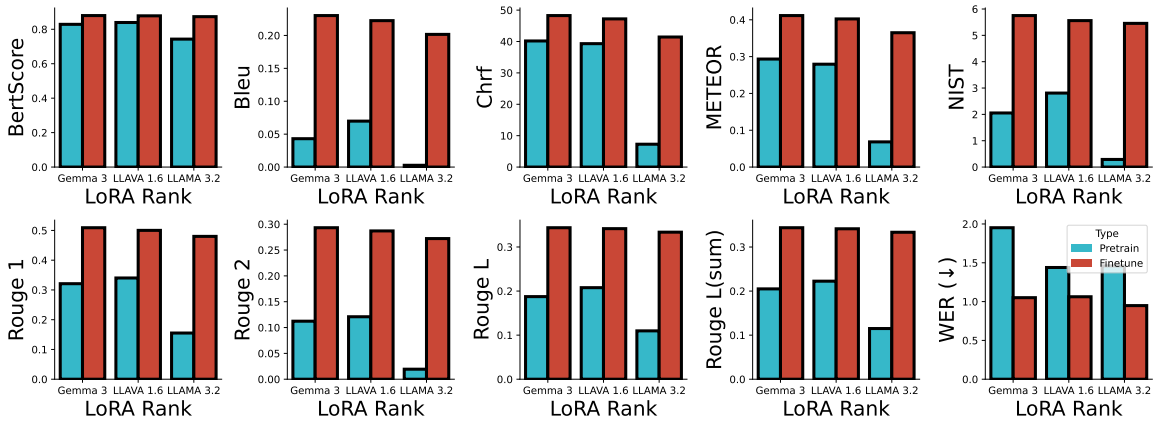

Figure 4: Performance of PatentVision with different base LVLMs compared to their pretrained versions.

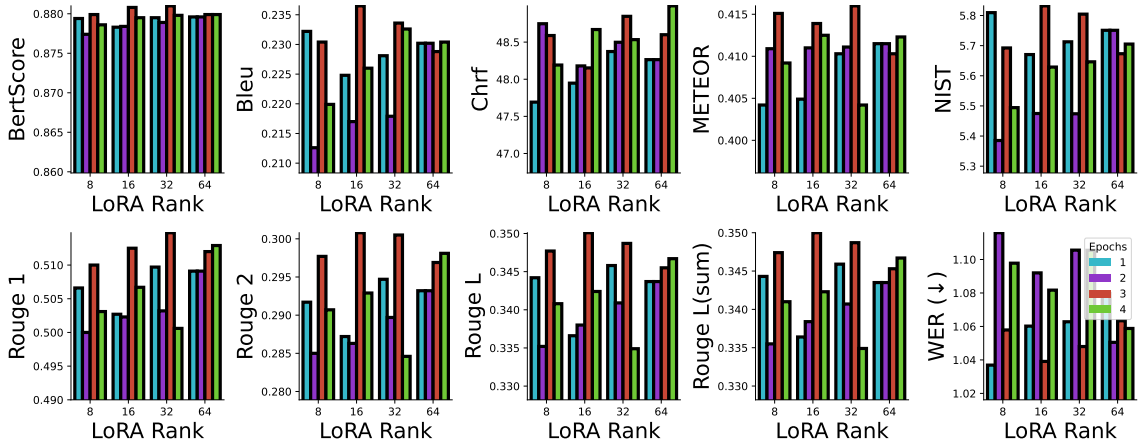

Figure 5: Performance of PatentVision with Gemma 3 as base model trained with varying epochs and LoRA ranks.

results than PatentFormer with image descriptions, demonstrating that PatentVision effectively extracts meaningful information directly from raw images.

5.4 Ablation study

Finally, we investigate the effects of varying LoRA ranks, image resolution, and number of epochs on the performance of PatentVision.

Impact of Lora Rank. Using a small LoRA rank may limit the model's capacity to acquire the domain-specific knowledge required for the patent writing task. Conversely, excessively large LoRA ranks can lead to convergence issues during training. Figure 3 shows PatentVision achieves better performance with mid-range LoRA ranks (e.g., 32, 64, and 128) across different base models. In contrast, training fails to converge with large LoRA ranks, e.g., 256 for LLAVA 1.6 and LLaMA 3.2.

More epochs with Gemma 3. As noted in the previous section, Gemma 3 outperforms LLAVA 1.6 and LLaMA 3.2 as the base model for PatentVision. To investigate the impact of training duration on generation quality, we train PatentVision using Gemma 3 across various LoRA ranks (8, 16, 32, and 64) with different numbers of training epochs. As shown in Figure 5, performance degrades when the model is trained for four epochs, indicating overfitting. In contrast, training for three epochs consistently yields superior results across different LoRA rank settings, indicating it as the optimal configuration for this task.

The effects of image resolution. Next, we examine the effect of image resolution on the quality of the generated specifications. Specifically, we conduct experiments using image resolutions of 256, 512, 1024, 2048, and 4096 pixels. As shown in Figure 7, higher image resolutions generally lead to improved generation quality. This trend suggests that increased resolution allows vision-language models to better capture and interpret fine-grained details within patent diagrams, which in turn enhances the overall specification generation.

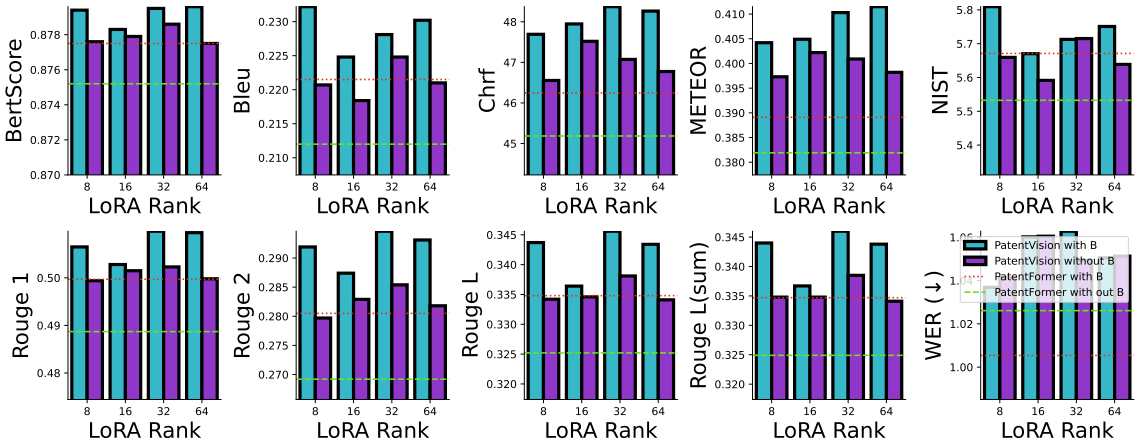

Figure 6: Performance of PatentVision with Gemma 3 as the base model trained with varying LoRA ranks on test sets with and without image descriptions (B).

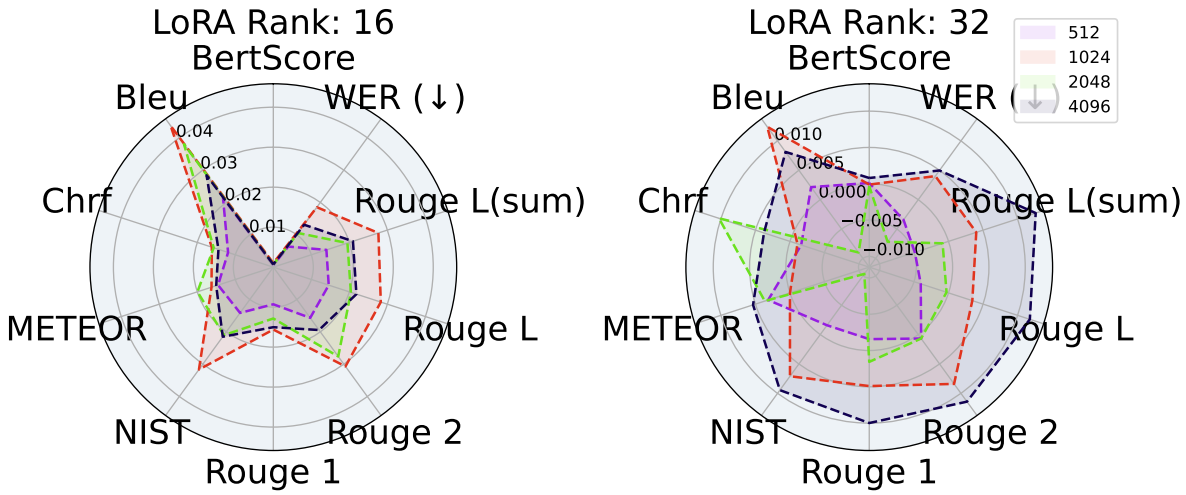

Figure 7: Performance improvement percentages compared to the performance of PatentVision using an image resolution of 256, for PatentVision with Gemma 3 as the base model across different metrics.

Chat functionality with Gemma 3. One of the key advancements of PatentVision over PatentFormer is its design as an interactive agent capable of accepting human instructions, images, and patent text as input, rather than relying solely on patent text. This capability enables the users to provide post-generation instructions, such as editing or refining the generated specification, thereby supporting iterative improvement. The interactive nature of PatentVision significantly enhances the potential quality of the output, as the users can guide the model to correct or elaborate on its own generation—something not possible with PatentFormer. We plan to incorporate full conversational functionality in the next version of PatentVision to further support this interactive workflow.

6 Conclusions

We proposed a novel method, PatentVision, to utilize diverse patent-related information, e.g., patent claims, drawings, and brief descriptions of the drawings, for generating patent specification. We leveraged large vision language models to generate specification by using both text and image modalities. Experimental evaluations affirmed the effectiveness and practical usefulness of our proposed methods.

Ethics Statement

Patents are legal documents, and the USPTO[4] recommends the practitioners to take extra care to verify the technical accuracy of the documents and compliance with 35 U.S.C. 112 when using AI drafting tools (Holman, 2024).

Limitation

The development and implementation of PatentVision have shown promising results, but there are limitations that need to be acknowledged. Specifically, PatentVision currently lacks a chat functionality, which restricts the interaction between the model and the user, hindering the ability to further improve the quality of the generated specification through iterative feedback and refinement. Furthermore, PatentVision generates specifications on a per-claim basis, and since a single patent often contains multiple claims, the processing time increases linearly with the number of claims, making the process costly for patents with numerous claims. These limitations highlight areas for future development and improvement to enhance the functionality and efficiency of PatentVision.

References

Dana Aubakirova, Kim Gerdes, and Lufei Liu. 2023. Patfig: Generating short and long captions for patent figures. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2843–2849.

Satanjeev Banerjee and Alon Lavie. 2005. 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, Ann Arbor, Michigan. Association for Computational Linguistics.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Dimitrios Christofidellis, Antonio Berrios Torres, Ashish Dave, Manuel Roveri, Kristin Schmidt, Sarath Swaminathan, Hans Vandierendonck, Dmitry Zubarev, and Matteo Manica. 2022. Pgt: a prompt based generative transformer for the patent domain. In *ICML 2022 Workshop on Knowledge Retrieval and Language Models*.

[4]https://www.federalregister.gov/documents/2024/04/11/2024-07629/guidance-on-use-of-artificial-intelligence-based-tools-in-practice-before-the-united-states-patent

George Doddington. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the Second International Conference on Human Language Technology Research*, HLT '02, page 138–145, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Shi Guoliang, Zhou Shu, Wang Yunfeng, Shi Chunjiang, and Liu Liang. 2023. Generating patent text abstracts based on improved multi-head attention mechanism. *Data Analysis and Knowledge Discovery*, 7(6):61–72.

Christopher M Holman. 2024. The us patent and trademark office's response to recent developments in artificial intelligence. *Biotechnology Law Report*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Lekang Jiang, Caiqi Zhang, Pascal A Scherz, and Stephan Goetz. 2024. Can large language models generate high-quality patent claims? *arXiv preprint arXiv:2406.19465*.

LEE Jieh-Sheng. 2022. The effectiveness of bidirectional generative patent language models. In *Legal Knowledge and Information Systems: JURIX 2022: The Thirty-fifth Annual Conference, Saarbrücken, Germany, 14-16 December 2022*, volume 362, page 194. IOS Press.

Jieh-Sheng Lee. 2020a. Controlling patent text generation by structural metadata. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 3241–3244.

Jieh-Sheng Lee. 2020b. Measuring and controlling text generation by semantic search. In *Companion Proceedings of the Web Conference 2020*, pages 269–273.

Jieh-Sheng Lee. 2020c. Patent transformer: A framework for personalized patent claim generation. In *CEUR Workshop Proceedings*, volume 2598. CEUR-WS.

Jieh-Sheng Lee. 2023. Evaluating generative patent language models. *World Patent Information*, 72:102173.

Jieh-Sheng Lee and Jieh Hsiang. 2020a. Patent claim generation by fine-tuning openai gpt-2. *World Patent Information*, 62:101983.

Jieh-Sheng Lee and Jieh Hsiang. 2020b. Patenttransformer-1.5: Measuring patent claim generation by span relevancy. In *New Frontiers in Artificial Intelligence*, pages 20–33, Cham. Springer International Publishing.

Jieh-Sheng Lee and Jieh Hsiang. 2020c. Prior art search and reranking for generated patent text. *arXiv preprint arXiv:2009.09132*.

Junyi Li, Tianyi Tang, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2024. Pre-trained language models for text generation: A survey. *ACM Computing Surveys*, 56(9):1–39.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Chin-Yew Lin and Franz Josef Och. 2004. ORANGE: a method for evaluating automatic evaluation metrics for machine translation. In *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, pages 501–507, Geneva, Switzerland. COLING.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024. Llava-next: Improved reasoning, ocr, and world knowledge.

Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. 2022. Biogpt: generative pre-trained transformer for biomedical text generation and mining. *Briefings in bioinformatics*, 23(6):bbac409.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. pages 311–318.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Cinthia M Souza, Magali RG Meireles, and Paulo EM Almeida. 2021. A comparative study of abstractive and extractive summarization techniques to label subgroups on patent dataset. *Scientometrics*, 126(1):135–156.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

Juanyan Wang, Sai Krishna Reddy Mudhiganti, and Manali Sharma. 2024. Patentformer: A novel method to automate the generation of patent applications. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1361–1380.

Sinong Wang, Belinda Z Li, Madian Khabsa, Han Fang, and Hao Ma. 2020. Linformer: Self-attention with linear complexity. *arXiv preprint arXiv:2006.04768*.

J.P. Woodard and J.T. Nelson. 1982. An information theoretic measure of speech recognition performance.

Chuhan Wu, Fangzhao Wu, Tao Qi, and Yongfeng Huang. 2021. Hi-transformer: Hierarchical interactive transformer for efficient and effective long document modeling. *arXiv preprint arXiv:2106.01040*.

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, et al. 2020. Big bird: Transformers for longer sequences. *Advances in neural information processing systems*, 33:17283–17297.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Changsheng Zhu, Xin Zheng, and Wenfang Feng. 2023. An automatic generation method of patent specification abstract based on" extraction-abstraction" model. In *2023 IEEE 3rd International Conference on Power, Electronics and Computer Applications (ICPECA)*, pages 196–200. IEEE.