# PATENTLMM: Large Multimodal Model for Generating Descriptions for Patent Figures

**Shreya Shukla**[1*], **Nakul Sharma**[1*], **Manish Gupta**[2], **Anand Mishra**[1]
[1]Indian Institute of Technology Jodhpur, India
[2]Microsoft, India
{shukla.12,sharma.86}@iitj.ac.in, gmanish@microsoft.com, mishra@iitj.ac.in

## Abstract

Writing comprehensive and accurate descriptions of technical drawings in patent documents is crucial to effective knowledge sharing and enabling the replication and protection of intellectual property. However, automation of this task has been largely overlooked by the research community. To this end, we introduce PATENTDESC-355K, a novel large-scale dataset containing ∼355K patent figures along with their brief and detailed textual descriptions extracted from more than 60K US patent documents. In addition, we propose PATENTLMM – a novel large multimodal model specifically tailored to generate high-quality descriptions of patent figures. Our proposed PATENTLMM comprises two key components: (i) PATENTMME, a specialized multimodal vision encoder that captures the unique structural elements of patent figures, and (ii) PATENTLLAMA, a domain-adapted version of LLaMA fine-tuned on a large collection of patents. Our extensive experiments demonstrate that training a vision encoder specifically designed for patent figures significantly boosts the performance, generating coherent descriptions compared to fine-tuning similar-sized off-the-shelf multimodal models. PATENTDESC-355K and PATENTLMM pave the way for automating the understanding of patent figures, enabling efficient knowledge sharing and faster drafting of patent documents.

**Project Page** —
https://vl2g.github.io/projects/PatentLMM/

## 1 Introduction

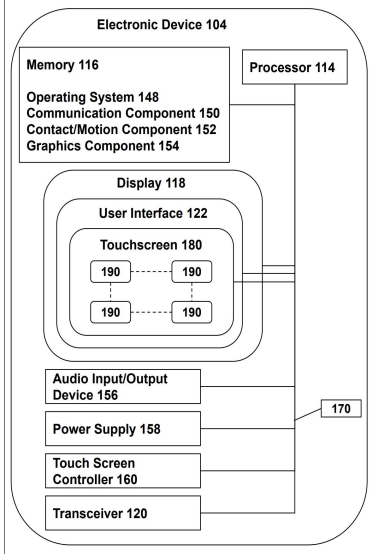
Patents are a cornerstone of intellectual property protection, granting inventors exclusive rights to their creations. Effective communication of these inventions is crucial for patent examiners, courts, and the technical community to appreciate the inventiveness of these inventions and assess their novelty. Patent documents rely heavily on figures and their corresponding textual descriptions to present technical details. Writing accurate descriptions of these figures is essential for an unambiguous understanding of the invention and its components and facilitates knowledge sharing within the technical community. Comprehensive descriptions also ensure that the invention is adequately protected against potential infringements by others. However, manually crafting such descriptions is time-consuming and laborious, hindering the efficiency of patent processing and analysis.

One of the major challenges for generating patent figure descriptions in an automated way is the lack of large-scale labeled datasets. Existing datasets, while invaluable for advancing research in natural and scientific figure captioning, do not adequately capture the nuances and complexities inherent to patent illustrations. To address this gap, we curate PATENTDESC-355K, a novel large-scale dataset containing ∼355K patent figures and their brief and detailed textual descriptions extracted from 60K+ patent documents. This dataset offers a rich and diverse collection of patent figures that span various technical domains, along with their corresponding descriptions, enabling the development and evaluation of models specifically tailored for this task.

Typically, patent figures are associated with brief and detailed descriptions. In our proposed PATENTDESC-355K dataset, we found that they span an average of ∼34 and ∼1680 tokens, respectively. Thus, unlike existing image captioning benchmarks, for example COCO (Lin et al. 2014), TextCaps (Sidorov et al. 2020) and NoCaps (Agrawal et al. 2019) where captions span an average of ∼12 tokens, the descriptive captioning of patent figures in our dataset is much more challenging. Moreover, unlike the natural scene images of the existing captioning datasets, patent figures are structured technical illustrations that adhere to a more standardized visual style for technical and legal documentation.

The emergence of Large Language Models (LLMs) and Large Multimodal Models (LMMs) has revolutionized almost every vision and language task. These models exhibit a remarkable ability to understand and generate coherent language across diverse domains. However, applying these models to the generation of patent descriptions presents unique challenges. The length of descriptions and the complexity inherent to patent diagrams underscore the need to focus on various elements of the figure, such as arrows, nodes, and text annotations. Further, contrary to dense document images, patent figures are sparse and comprise several elements like text, nodes, node labels (a number associated with nodes in the patent figure), figure numbers, and arrows in different styles, i.e., uni-direction and bidirectional, solid, and dotted, among others. Please refer to Fig. 3 in the Appendix for an overview of these elements.

[*]These authors contributed equally.




| Generated Brief and Detailed Descriptions | Ground Truth Brief and Detailed Descriptions |
|---|---|
| FIG. 1 illustrates a mobile smartphone in accordance with an aspect of the invention. | FIG. 1 shows the details of an exemplary electronic device in accordance with aspects of the invention. |
| Referring now to FIG. 1, an exemplary electronic device 104 is illustrated. It should be understood that the electronic device 104 may be any one of a number of electronic devices including, but not limited to, a cell phone, a smart phone, a personal digital assistant, a laptop computer, a desktop computer, a netbook computer, a server computer, etc. The electronic device 104 includes a processor 114, a memory 116, a display 118, a user interface 122, a transceiver 120, and an optional touch screen 160. The processor 114 is configured to execute instructions received from the memory 116, and is, for example, a general purpose processor, a field programmable gate array, or any other suitable processor. The processor 114 is configured to execute instructions received from the memory 116, including instructions for displaying a user interface on the display 118. The processor 114 may also be configured to execute instructions received from the transceiver 120, including instructions for receiving data from the transceiver 120 and instructions for transmitting data from the transceiver 120.The memory 116 is configured to store data received from the processor 114 and instructions to be executed by the processor 114. The memory 116 may be, for example, a random access memory device, a solid state memory device, or any other suitable …. | FIG. 1 shows the details of an exemplary electronic device in accordance with aspects of the invention. The electronic device 104 includes a processor 114, memory 116, display 118, user interface 122, and the like. The processor 114 may be a central processing unit configured to execute instructions including instructions related to software programs. The display 118 may be a liquid crystal display having a backlight to illuminate the various color liquid crystals to provide a colorful display. The user interface 122 may be any type of physical input having buttons and further may be implemented as a touchscreen 180.The electronic device 104 may further include in the memory 116, an operating system 148, a communication component 150, a contact/motion component 152, a graphics component 154, and the like. The operating system 148 together with the various components providing software functionality for each of the components of the electronic device 104.The memory 116 may include a high-speed random-access memory. Also, the memory 116 may be a non-volatile memory, such as magnetic fixed disk storage, flash memory or the like. These various components may be connected through various communication lines including a data bus 170.Additionally, the electronic device 104 may include an audio input/output device 156. The audio … |

Figure 1: An example of generated and ground truth brief and detailed descriptions using our proposed PATENTLMM.

As can be seen in Fig. 1, the detailed description of patent figures heavily makes use of these elements to convey the semantics of the figure. Given this dramatic difference between captions of natural scenes versus patent figures, it was anticipated that recent image captioning methods (Li et al. 2022; Wang et al. 2022a,b) and multimodal LLMs (Ye et al. 2023b; Liu et al. 2024a; Zhu et al. 2024) would perform poorly for our task in a zero-shot setting. Surprisingly, these approaches demonstrated suboptimal performance even after fine-tuning on our dataset. These unique properties of patent figures require specialized system design to ensure the accurate and concise generation of descriptions without introducing hallucinations or irrelevant details.

In this paper, we propose PATENTLMM – a novel model to generate descriptions of patent figures. The model contains two important components: PATENTMME and PATENTLLAMA. PATENTMME is a specialized multimodal vision encoder for patent figures, trained using masked language modeling loss, along with two other novel loss functions focused on learning structure from sparse patent figures. PATENTLLAMA is a domain-adapted version of LLaMA fine-tuned on a large collection of patent text from the Harvard USPTO Dataset (HUPD) (Suzgun et al. 2024). PATENTLMM combines the PATENTMME encoder and the PATENTLLAMA using a projection layer.

The major contributions of our work are as follows. (i) We present a large-scale dataset of $\sim$355K patent figures and their brief and detailed descriptions. (ii) We propose a novel multimodal model PATENTLMM, comprising a patent domain-specialized vision encoder trained using objectives specifically tailored to capture the structure of patent documents and an LLM fine-tuned on patent data. (iii) We extensively benchmark existing captioning models and multimodal LLMs and show that our proposed approach surpasses their best performance on the average BLEU metric by 10.22% and 4.43% on an absolute scale for generating brief and detailed descriptions, respectively.

## 2 Related Work

**Image Captioning in Pre-LMMs era:** The patent figure description task is broadly similar to the image captioning task, which has been an active research area in the last decade. Some representative early work on image captioning includes the combination of a CNN encoder with an LSTM decoder (Vinyals et al. 2015), a multimodal RNN architecture that uses local and global image features (Andreas et al. 2016), an adaptive attention model (Lu et al. 2017), and a bottom-up and top-down attention model (Anderson et al. 2018). Recent works have also focused on improving caption diversity (Shetty, Roumeliotis, and Laaksonen 2017), novel object captioning (Lu et al. 2018), and incorporating external knowledge (Gu et al. 2019). As discussed in the previous section, our task differs significantly from these previous efforts on image captioning in terms of the length of descriptions and the structure of patent figures.

**Describing Scientific Figures:** Patent figures are a specific form of scientific illustrations. Although previous work on generating descriptions of patent figures has been sparse, ample research has been done to caption scientific figures. Chen et al. (2019, 2020) create and leverage FigCAP and adapt an LSTM-based model (Hochreiter and Schmidhuber 1997) for captioning. Recently, Hsu, Giles, and Huang (2021) collected the SciCap dataset from articles published on arXivIn (Yang et al. 2023), the authors augment the SciCap dataset with additional information such as OCR text from figures and referring sentences from the text to curate SciCap+, and demonstrate the performance boost achieved by incorporating extra information. Kantharaj et al.

(2022) and Tang, Boggust, and Satyanarayan (2023) address the problem of captioning various visualization charts of data. Certain works go beyond natural language descriptions to generate code, particularly for flowcharts. For example, Shukla et al. (2023) and Liu et al. (2022) specifically address the generation of code from flow chart images.A parallel work PatFig (Aubakirova, Gerdes, and Liu 2023) scrapes a similar dataset as ours with 17K training samples and 2K test samples, and demonstrates the performance of MiniGPT-4 (Zhu et al. 2024) in the proposed dataset. In this work, we contribute a $\sim$20$\times$ larger dataset and propose a novel model, PATENTLMM, which is almost twice as effective as MiniGPT-4 in BLEU-4 for PATENTDESC-355K.

**Large Multimodal Models:** Recent work in the multimodal (vision and language) community has focused on leveraging the world knowledge implicitly encoded in large language models for multimodal tasks such as visual question answering and image captioning (Zhu et al. 2024; Li et al. 2023; Liu et al. 2024a; Achiam et al. 2023; Ye et al. 2023b, 2024; Wang et al. 2022b; Team et al. 2023; Alayrac et al. 2022), visual grounding (Ye et al. 2024; Zhu et al. 2024; Team et al. 2023; Achiam et al. 2023) and image-text matching (Li et al. 2022, 2023). This is achieved by feeding an image representation as input along with the prompt to the language model and modeling the output using the language modeling objective. Recent advances include Flamingo (Alayrac et al. 2022), which inserts trainable gated cross-attention layers into a pretrained LLM (Hoffmann et al. 2022). BLIP-2 (Li et al. 2023) leverages pre-trained ViT (Dosovitskiy et al. 2021) and LLaMA (Touvron et al. 2023), combined with QFormer, to translate image embeddings into LLM prompt embeddings. MiniGPT-4 (Zhu et al. 2024) builds upon pretrained BLIP-2 and finetunes an additional linear layer to project queries into the LLM on a curated dataset. In contrast, LLaVA-1.5 (Liu et al. 2024a) proposes a relatively simple and effective two-stage approach. In addition, document-specific LLMs such as LayoutLLM (Luo et al. 2024), UReader (Ye et al. 2023a) and TextMonkey (Liu et al. 2024b) have shown impressive performance on Document VQA task. We compare with several of these models and show that these models do not perform competitively for the task of generating descriptions from patent figures.

## 3 PATENTDESC-355K: A Novel Dataset of Patent Figures with Descriptions

We introduce PATENTDESC-355K – a novel large-scale dataset tailored for generating descriptions for patent figures. Our proposed dataset comprises 355K patent figures sourced from Google Patents[1], with each image accompanied by its brief and detailed descriptions extracted from the corresponding patent documents. The dataset is available for download on our project website: https://vl2g.github.io/projects/PatentLMM/. Fig. 1 visualizes a ⟨patent figure, brief description, detailed description⟩ triplet from our dataset. With our primary focus on US patents published after 2004, our dataset spans over 60K patents from assignees like Amazon, Microsoft, LinkedIn, Google, Yahoo, etc. To assess the quality of the dataset, we manually evaluated a random set of 100 patent figures with their brief and detailed descriptions and computed the sentence-level precision and recall of the extracted descriptions against the ground-truth descriptions. For brief descriptions, both precision and recall scores were 100%. For detailed descriptions, precision and recall were 90. 81% and 91. 96%, respectively. More details on data set curation, preprocessing, description extraction, and quality assessment are provided in Appendix A.

[1]https://patents.google.com

| | Train | Validation | Test |
|---|---|---|---|
| Number of Images | 320,717 | 17,286 | 17,336 |
| Avg. number of tokens in brief descriptions | 34.37 | 34.28 | 34.30 |
| Avg. number of tokens in detailed descriptions | 1,677.85 | 1,676.71 | 1,697.16 |
| Number of Unique Patents | 50,448 | 8,027 | 7,964 |
| Avg. number of images per patent | 6.36 | 2.15 | 2.18 |

Table 1: PATENTDESC-355K: Dataset Statistics.

**Dataset Analysis**: Table 1 presents detailed statistics of the 355K image-description triplets in our dataset. During the creation of training, validation and test set splits, we ensure absolute exclusivity between patents in the train set and those in the combined validation and test sets, to enable robust out-of-sample evaluation. To achieve this, we randomly sampled $\sim$ 12.6K patents from $\sim$ 60K, representing $\sim$ 82.5K images. From this isolated subset of images, we sample $\sim$17K images each for the val and test set, and discard the remaining images. This sampling technique also helps maintain the diversity within the validation and test sets, thereby providing a fair and representative evaluation. Our detailed descriptions span $\sim$1.7K tokens on average, which is much larger compared to an average token length for popular image captioning benchmarks (Lin et al. 2014; Chen et al. 2015; Sidorov et al. 2020).

## 4 Methodology

Our approach is inspired by the recent success of large multimodal models like MiniGPT-4 (Zhu et al. 2024) and LLaVA (Liu et al. 2023, 2024a), which have demonstrated state-of-the-art performance on several benchmarks by effectively aligning visual and textual modalities. We introduce PATENTLMM, which combines our domain-adapted version of the LLaMA language model, namely PATENTLLAMA, with our novel visual encoder specialized for patent figures, namely PATENTMME. In this section, we describe the architectures of PATENTMME and PATENTLLAMA, and the overall framework of PATENTLMM.

### 4.1 PATENTMME: Encoder for Patent Figures

The Vision Transformer (ViT) (Dosovitskiy et al. 2021), commonly used as a vision encoder in existing image captioning frameworks, is typically pre-trained on natural scene images, which are fundamentally different from patent figures. A better suited encoder is perhaps LayoutLM (Xu et al. 2020, 2021; Huang et al. 2022) which has shown impressive performance in document image understanding tasks. However, patent figures have a sparse layout compared to dense document images and are characterized by specific structured visual syntax. Unlike document images, patent figures

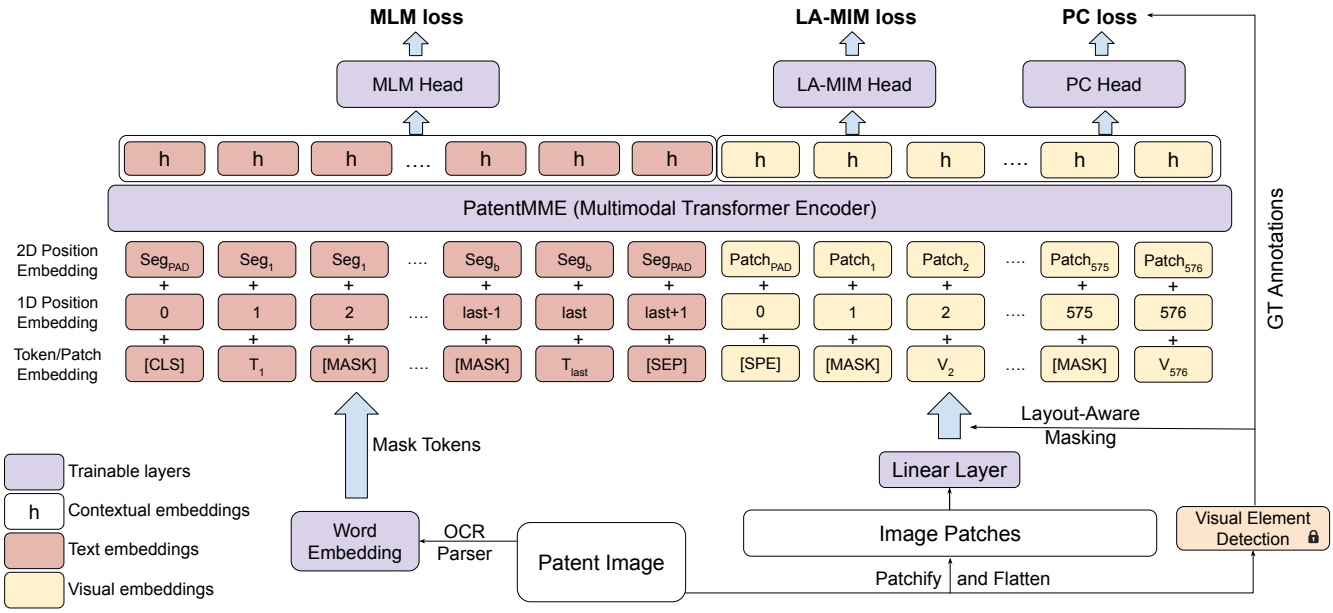



Figure 2: PATENTMME Architecture. We jointly process OCR tokens and visual embeddings to produce multimodal context-aware embeddings. These contextual embeddings are optimized using our proposed MLM, LA-MIM and PC objectives.

comprise labeled nodes interconnected with arrows and accompanied by textual elements. The semantic relationship between these diagrammatic constituents is paramount for decoding the inventive concepts and technical specifications elucidated within the patent figures. We, therefore, build on the existing document image understanding capabilities of LayoutLMv3 (Huang et al. 2022) and pre-train it with novel objectives, specifically tailored to capture the structural information of patent figures.

**Architecture:** The proposed PATENTMME shares its architecture with LayoutLMv3 (Huang et al. 2022) and is a multi-modal transformer model that processes image, text, and document layout information jointly. The overall architecture of the model is illustrated in Fig. 2. Given an input patent figure $I$, the OCR text is extracted using off-the-shelf Tesseract OCR engine (Kay 2007). The image is then down-scaled to $H \times W$ and split into non-overlapping patches of $p$ dimensions each, resulting in $M = HW/p^2$ image patches. The OCR extracted text is tokenized using the BPE tokenizer (Shibata et al. 1999) and represented using a learnable embedding matrix. Following (Huang et al. 2022), learnable 1D-position embeddings and 2D segment-level layout-position embeddings are added to the word embeddings, resulting in the final text embeddings. The image embeddings are created by linear projection of flattened image patches and combining them with learnable 1D position embeddings and 2D spatial embeddings. We use images of size $I \in \mathbb{R}^{3\times384\times384}$, i.e., $H = W = 384$. With $p = 16$ this results in $M = 576$ patches. The higher resolution helps preserve intricate structural details of patent figures, such as node labels and arrows.

**Pre-training data and annotations:** To enable large-scale in-domain pre-training of PATENTMME, we crawled a set of 900K+ patent figures corresponding to the patent IDs from the Harvard USPTO Patent Dataset (HUPD) (Suzgun et al. 2024). For a fair evaluation, appropriate care has been taken to avoid any overlap of the sample with the validation and testing split of our dataset.

For robust patent-figures'-specific pretraining, we define loss functions that leverage patent diagram specific elements like nodes, node labels, figure labels, text and arrows. To extract such elements, we train a Faster-RCNN (Ren et al. 2015) based visual element detection network on 350 manually annotated patent figures, sampled randomly from our training data. The trained model is then used to infer elements from all training images, which is used to provide weak ground-truth labels during PATENTMME training. We show inference samples of this model in Appendix B.

**Pre-training Loss Formulations:** To enhance the vision encoder's capability in capturing fine-grained structural details of patent figures, we pre-train PATENTMME using novel layout-aware masked image modeling (LAMIM) and image patch classification (PC) objectives, along with the established masked language modeling (MLM) loss. We describe these losses in the following text.

**Notation:** We use $R$ and $T$ to denote the set of image patches (regions) and OCR tokens, respectively. Further, $X_m$ and $X_{um}$ denote the masked and unmasked parts of the modality $X$. The probability distribution generated by our PATENTMME model and the set of categories of visual elements that can be detected by our detection network by $p_\theta$ and $C$, respectively.

(i) **Masked Language Modeling (MLM)**. Similar to LayoutLMv3 (Huang et al. 2022), we randomly mask 30% of the OCR text tokens and optimize the model to predict the masked tokens, encouraging it to learn patent-specific tex-

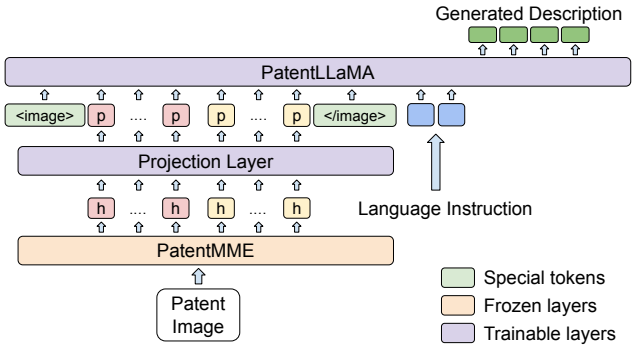



Figure 3: PATENTLMM Architecture. Language Instruction is a fixed prompt guiding the model to generate either brief or detailed descriptions.

tual semantics. This loss is computed as follows:

$$\mathcal{L}_{MLM}(\theta) = -\sum_{i \in T_m} \log p_\theta(t_i \mid R_{um}, T_{um}), \qquad (1)$$

where $t_i$ denotes the correct masked text tokens.

(ii) **Layout-Aware Masked Image Modeling (LAMIM)**. We utilize masked image modeling to learn visual representations by randomly masking 40% of the image patches. Since the patent figures are more sparse compared to dense document images, we mask only the image patches that contain atleast one of the following five elements: nodes, node labels, figure labels, text and arrows. This strategy helps to avoid masking blank regions in the patent figures, and hence learn robust visual representations. Our formulation of the LAMIM objective is similar to BEiT (Bao et al. 2022) and therefore requires a discrete image tokenizer. We choose OCR-VQGAN (Rodriguez et al. 2023) since its tokenized image representation is capable of handling textual information better than competing works dVAE (Ramesh et al. 2021) and VQGAN (Esser, Rombach, and Ommer 2021). We compute this loss as follows:

$$\mathcal{L}_{MIM}(\theta) = -\sum_{i \in R_m} \log p_\theta(r_i \mid R_{um}, T_{um}), \qquad (2)$$

where $p_i$ denotes the correct masked image patches.

(iii) **Patch Classification (PC).** In this multi-label binary classification objective, we classify each of the $M$ image patches into one or more of the following five categories: *node*, *node label*, *figure label*, *text*, and *arrows*. This objective which is mathematically computed as follows, helps the model learn discriminative representations for different visual elements in patent figures.

$$\mathcal{L}_{PC} = -\frac{1}{M}\sum_{i=1}^{M}\sum_{j=1}^{|C|}\left[y_{ij}\log(\hat{y}_{ij}) + (1-y_{ij})\log(1-\hat{y}_{ij})\right], \qquad (3)$$

where $\hat{y}_{ij}$ denotes the probability of patch $i$ belonging to class $j$, and $y_{ij}$ is the binary ground truth label obtained using the visual element detector.

### 4.2 PATENTLLAMA: Description Generator

PATENTLLAMA is a domain-adapted version of the LLaMA-2 7B model for the patent domain. We continue to pre-train the LLaMA-2 7B model using LoRA (Hu et al. 2022) adapters, on the descriptions from HUPD patent dataset (Suzgun et al. 2024), to bias the model to generate the language inherent to patent documents. To avoid any train-test leakage, we ensure that we use the HUPD dataset after removing patent documents corresponding to the validation and test splits of our PATENTDESC-355K dataset.

### 4.3 PATENTLMM

Inspired by recent multimodal LLM studies like MiniGPT-4 (Zhu et al. 2024) and LLaVA (Liu et al. 2023, 2024a), we integrate PATENTMME and PATENTLLAMA through a single MLP network to exploit their pre-trained representations. The detailed architecture for PATENTLMM is illustrated in Fig. 3. Given a patent figure, we first obtain its layout-aware text and visual representations from frozen PATENTMME. These representations are projected into the input embedding space of PATENTLLAMA using a projection MLP, and the PATENTLLAMA is finetuned to maximize the likelihood of the corresponding description conditioned on these projected representations.

## 5 Experiments

### 5.1 Experimental Setup

**PATENTMME:** PATENTMME is initialized with LayoutLMv3-Large to inherit its document understanding capabilities. For each of the three losses discussed in Section 4, the text and image embeddings obtained from PATENTMME are projected through separate MLPs (loss heads) before the loss is calculated. Since the network weights already have a good initialization, to prevent major changes in weights of the multimodal transformer, we adopt two-step training. During Step-1, the weights of the multimodal transformer remain frozen and only the loss heads are trained for 1 epoch with a higher learning rate of 1e-3 and 1K warm-up steps to learn good initialization. During Step 2, the entire model is trained end-to-end for 8 epochs with a lower learning rate of 5e-5 and with 10K warm-up steps. The PATENTMME model is trained on 8×V100 GPUs, with an effective batch size of 64 and Adam (Kingma and Ba 2014) optimizer.

**PATENTLMM:** Following the standard practice (Liu et al. 2024a), we train our PATENTLMM model in two stages. To align the patent figure representations obtained from PATENTMME with the input latent space of PATENTLLAMA, we train only the projection layer in the first stage, keeping all other parameters frozen. During stage 2, we add LoRA adapters to all the linear layers of the PatentLLaMA module, except for the language modeling head, whose weights remain frozen. The weights of PATENTMME are kept frozen throughout. We train our PATENTLMM with an effective batch size of 192 on 3×A100 GPUs (40 GB). Stage 1 training progresses at a higher learning rate of 1e-3, and stage 2 training takes place

| Setup | Method | # Parameters | B-2 | B-4 | Avg. B | R-1 | R-2 | R-L | M | B-2 | B-4 | Avg. B | R-1 | R-2 | R-L | M |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | Brief | | | | | | | Detailed | | | | | | |
| Zero-shot | BLIP-2 | 2.7B | 1.01 | 0.03 | 1.62 | 15.43 | 1.70 | 12.47 | 6.72 | 0.00 | 0.00 | 0.01 | 3.24 | 0.49 | 2.84 | 1.00 |
| | TextMonkey | 9.8B | 0.91 | 0.11 | 1.12 | 13.00 | 4.60 | 12.16 | 7.14 | 0.10 | 0.03 | 0.10 | 6.18 | 2.38 | 4.83 | 2.64 |
| | PEGASUS | 568M | 3.18 | 0.13 | 4.20 | 14.68 | 2.33 | 11.46 | 13.24 | 0.86 | 0.04 | 1.12 | 12.26 | 1.96 | 9.47 | 5.18 |
| | mPLUG-owl2 | 7.5B | 3.64 | 0.36 | 4.47 | 21.47 | 5.10 | 19.41 | 13.07 | 3.65 | 0.49 | 3.40 | 23.75 | 5.39 | 14.85 | 11.83 |
| | UReader | 7.2B | 3.54 | 0.35 | 4.50 | 20.90 | 4.56 | 17.85 | 13.45 | 0.05 | 0.01 | 0.06 | 5.15 | 1.54 | 4.49 | 2.04 |
| | LLaVA-1.5 | 7.4B | 4.52 | 0.24 | 4.71 | 17.59 | 3.27 | 14.63 | 15.74 | 3.65 | 0.37 | 3.36 | 23.75 | 4.63 | 14.71 | 11.69 |
| | GPT-4V | Unknown | 20.74 | 8.56 | 18.68 | 36.07 | 15.65 | 31.89 | 32.88 | 19.61 | 6.05 | 18.26 | 39.95 | 12.14 | 20.16 | 27.31 |
| Finetuned | Pegasus | 568M | 2.44 | 0.14 | 4.03 | 13.86 | 1.55 | 11.52 | 11.62 | 5.80 | 0.41 | 6.33 | 19.28 | 2.24 | 15.27 | 12.11 |
| | GIT | 681M | 26.95 | 15.33 | 24.78 | 45.28 | 27.17 | 42.29 | 44.27 | 6.33 | 1.18 | 6.23 | 13.66 | 3.17 | 10.87 | 10.68 |
| | BLIP | 252M | 24.62 | 12.52 | 22.40 | 42.59 | 23.78 | 39.16 | 42.84 | 5.45 | 1.05 | 5.31 | 12.42 | 2.89 | 9.46 | 9.55 |
| | MiniGPT-4 | 7.8B (3.2M) | 30.57 | 17.96 | 28.13 | 43.53 | 25.33 | 40.35 | 43.03 | 11.01 | 2.81 | 10.26 | 28.91 | 6.23 | 15.67 | 16.65 |
| | OFA | 472M | 33.01 | 21.76 | 31.24 | 54.26 | 37.94 | 51.47 | 44.89 | 15.76 | 7.23 | 14.93 | 33.20 | 13.70 | 22.89 | 21.17 |
| | LLaVA-1.5 | 7.4B (341M) | 36.64 | 25.00 | 34.37 | 48.92 | 32.01 | 45.87 | 48.23 | 20.90 | 11.12 | 19.81 | 36.86 | 15.68 | 24.48 | 24.71 |
| | **PATENTLMM** | 7.4B (341M) | **46.40** | **36.66** | **44.59** | **56.68** | **42.63** | **54.18** | **56.44** | **25.42** | **15.02** | **24.24** | **40.70** | **19.27** | **27.54** | **28.39** |

Table 2: Quantitative results on PATENTDESC-355K (test set) for brief and detailed description generation (B=BLEU, R=ROUGE, M=METEOR). Number in parenthesis under **# Parameters** column denote number of trainable parameters.

at a learning rate of 2e-4 with a cosine schedule, for 12K steps using Adam optimizer. We train separate LMMs for brief and detailed descriptions.

Overall, training PATENTLMM is a three-phase process. Firstly, we train the PATENTMME encoder in a semi-supervised fashion by leveraging a vast amount of patent figures corresponding to patents in the HUPD dataset. Secondly, we domain-adapt the LLaMA-2 7B model on the HUPD patent text data to create PATENTLLAMA. Lastly, we integrate PATENTMME and PATENTLLAMA to create PATENTLMM, and train it following the two-stage process.

## 5.2 Baselines

We benchmark the performance of various baselines on our proposed PATENTDESC-355K dataset in the zero-shot and fine-tuned setup. We benchmark the text-only baseline Pegasus (Zhang et al. 2020) by generating patent figure descriptions from OCR tokens extracted from patent figures. For image captioning baselines, we study the state-of-the-art models GIT (Wang et al. 2022a), BLIP (Li et al. 2022) and OFA (Wang et al. 2022b). We further compare our method with recent multimodal LLMs such as UReader (Ye et al. 2023a), TextMonkey (Liu et al. 2024b), mPLUG-owl2 (Ye et al. 2024), BLIP-2 (Li et al. 2023), MiniGPT-4 (Zhu et al. 2024), LLaVA-1.5 (Liu et al. 2024a) and the closed GPT-4V model (Achiam et al. 2023). GPT-4V prompt is listed in Appendix C.3.

To measure the description generation performance of these models, we use standard image captioning metrics such as BLEU (Papineni et al. 2002), ROUGE (Lin 2004) and METEOR (Banerjee and Lavie 2005). Higher values for all the scores are desired. A detailed description of these metrics is provided in Appendix C.1.

## 5.3 Results and Discussion

The quantitative performance comparison for the brief and detailed description generation task is reported in Table 2. In the zero-shot setting, GPT-4V demonstrated superior performance among baselines across all metrics, significantly outperforming other baselines owing to its large scale and the diverse data it has seen during its pre-training. The poor zero-shot performance of other baselines highlights the gap in their pre-training data and the nature of patent figures and descriptions. The fine-tuned models outperform their zero-shot counterparts, highlighting the importance of task-specific training for these models. MiniGPT-4 and LLaVA-1.5 utilize a frozen pre-trained ViT trained on web-scale natural images, which results in suboptimal representation of patent figures. Similarly, OFA also enforces these priors by utilizing a pre-trained discrete image tokenizer. On the other hand, PATENTLMM gives a boost of $\sim 8\%$ across all metrics, signifying the importance of better domain knowledge embedded in it through the proposed PATENTMME pretraining and PatentLLaMA.

Similar to brief description generation, GPT-4V outperformed all other baselines for the detailed description generation task in the zero-shot setting. We observe that majority of the baselines struggle with performance in the zero-shot setup. In the fine-tuned setting, our PATENTLMM maintained its superior performance, achieving the highest scores across all metrics. This consistent top performance for both brief and detailed descriptions suggests the efficacy of our proposed approach for the task of generating descriptions from patent figures. The overall lower scores for detailed descriptions can be attributed to their comprehensiveness, complexity, and length, requiring models to capture and generate more nuanced and detailed information.

**Ablations:** We perform the following three ablation studies to quantify the impact of different components of our proposed PATENTLMM model:

**(i) PATENTMME Pre-training objectives**: Table 3 shows the ablation results with combinations of pre-training objectives for the brief description generation. We observe that using a combination of MLM and LAMIM leads to better results compared to the pre-trained LayoutLMv3. Further, the PC loss also improves the performance of the model, when pre-trained with HUPD images data. A similar ablation for

| Pre-training | B-2 | B-4 | Avg. B | R-1 | R-2 | R-L | M |
|---|---|---|---|---|---|---|---|
| Pretrained LayoutLMv3 | 42.81 | 32.50 | 40.86 | 53.68 | 38.88 | 51.07 | 53.34 |
| w/ MLM + LAMIM | 45.24 | 35.33 | 43.39 | 55.69 | 41.38 | 53.20 | 55.34 |
| w/ MLM+LAMIM+PC | **46.39** | **36.65** | **44.59** | **56.68** | **42.62** | **54.18** | **56.44** |

Table 3: Ablation study to quantify the impact of pretraining objectives of PATENTMME on the overall performance of PATENTLMM on brief descriptions generation task. All models are trained with PATENTLLAMA.

| OCR in training? | OCR in Inference? | B-2 | B-4 | Avg. B | R-1 | R-2 | R-L | M |
|---|---|---|---|---|---|---|---|---|
| No | No | 30.32 | 19.17 | 28.30 | 41.61 | 25.21 | 38.95 | 41.46 |
| Yes | No | 11.51 | 2.77 | 9.83 | 24.38 | 7.92 | 21.68 | 22.52 |
| Yes | Yes | **46.40** | **36.66** | **44.59** | **56.68** | **42.63** | **54.18** | **56.44** |

Table 4: Ablation study to quantify the importance of OCR tokens on the overall performance of PATENTLMM on brief descriptions generation task.

detailed descriptions is reported in Appendix C.2.

**(ii) Importance of OCR tokens**: In this ablation, we study whether avoiding passing OCR tokens to PATENTLMM causes any drop in the performance of brief description generation. We experiment with two ablations: (1) OCR tokens are used for PATENTMME pretraining but not for PATENTLMM training, and (2) OCR tokens are used for PATENTMME pretraining and for PATENTLMM training but not at inference time. Table 4 shows that it is important to use OCR tokens in the entire pipeline for the best results.

**(iii) PatentLMM Training:** We report an additional ablation study in Appendix C.2 to quantify the advantage of using PatentLLaMA against the pre-trained LLaMA model.

**Qualitative Analysis:** Fig. 1 shows an example brief and detailed description generated by PATENTLMM for a test sample. The generated brief description, more specifically, terms the electronic device shown in the image as a mobile smartphone. The generated detailed description provides a comprehensive overview of the electronic device 104, its components, and their functions. It covers most of the key elements mentioned in the ground truth, including the processor 114, the memory 116, the display 118, and the user interface 122. However, there are some omissions, like Graphics Component 154 and Communication Component 150. More case studies are provided in Appendix D.

**Error analysis**: We perform a thorough manual error analysis on a set of 50 samples drawn from our test set to identify some prominent errors in the descriptions generated by our PATENTLMM model. We identify five main error categories as follows. (i) Hallucination in figure labeling occurs in 3 brief and 3 detailed descriptions. (ii) Hallucination in 4 brief descriptions and 7 detailed descriptions was due to little or no OCR detectable text in the figures. (iii) Incorrect association of node labels occurs when the wiggly arrows connecting node labels to respective nodes are misinterpreted or ignored due to downsampling of the image before being passed to PATENTMME. This was observed in 10 detailed descriptions. (iv) A similar misinterpretation due to downsampling is often the cause of hallucinated node labels in 12 detailed descriptions. (v) Cross-figure references in the descriptions establish the interconnection between various aspects and provide a complete picture of the presented technical invention. The figures may be related hierarchically (systems vs components), sequentially (steps of a process), different views (top-bottom-left-right), or in other ways. Since we train PATENTLMM to generate descriptions for individual patent figures, our model hallucinates the cross-figure references for 2 brief and 5 detailed descriptions. Qualitative examples are presented in Appendix D.2.

| Description | Method | Rel. | Acc. | Compl. | Coh. | Fluency | Cover. |
|---|---|---|---|---|---|---|---|
| Brief | LLaVA-1.5 | 1.38 | 1.06 | 1.01 | 1.85 | 1.98 | 0.98 |
| | Ours | 1.44 | 1.18 | 1.17 | 1.91 | 2.00 | 1.15 |
| Detailed | LLaVA-1.5 | 0.75 | 0.75 | 0.73 | 1.07 | 1.69 | 0.71 |
| | Ours | 0.90 | 0.78 | 0.76 | 1.15 | 1.85 | 0.75 |

Table 5: GPT-4V evaluation on a set of 1K samples.

**GPT-4V Evaluation Results**: Apart from small-scale manual error analysis, we utilize the GPT-4V model to qualitatively evaluate the performance of LLaVA-1.5 and our proposed PATENTLMM model on the brief and detailed description generation task for a set of 1000 samples. We input the GPT-4V model with the patent figure, the ground truth description and the description generated using these models, along with the special instruction prompt. The instruction prompt instructs the GPT-4V model to rate the generated description on the following criterion: Relevance, Accuracy, Completeness (with respect to input image), Fluency and Coverage (with respect to input image and ground truth description) on an integer scale of 0 to 2. To mitigate randomness in scores, we set the temperature parameter to 0 for the GPT-4V model and created five versions of the instruction prompt. The scores obtained from each of the prompts for each criterion are then averaged. Table 5 shows that our system generates high-quality results.

## 6 Conclusion and Future Work

Our work addresses the existing gap in the automated generation of patent figure descriptions by introducing PATENTDESC-355K, a comprehensive dataset of patent figures and their corresponding brief and detailed descriptions. We further proposed PATENTLMM, a large multimodal model comprising a domain-specialized image encoder PATENTMME and a domain-adapted patentLLaMA model for generating brief and detailed descriptions from patent figures. Extensive experiments demonstrated that our proposed PATENTLMM outperforms competent baselines by significant margins. Future research in this direction can explore experiments with patents in multiple languages, patent document-level reasoning to allow for cross-figure references while generating descriptions, incorporating external knowledge bases from technical domains to improve the performance of detailed description generation, and generation of grounded descriptions.

## Acknowledgements

This work was supported by the Microsoft Academic Partnership Grant (MAPG) 2023.

## References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.

Agrawal, H.; Desai, K.; Wang, Y.; Chen, X.; Jain, R.; Johnson, M.; Batra, D.; Parikh, D.; Lee, S.; and Anderson, P. 2019. Nocaps: Novel object captioning at scale. In *ICCV*.

Alayrac, J.-B.; Donahue, J.; Luc, P.; Miech, A.; Barr, I.; Hasson, Y.; Lenc, K.; Mensch, A.; Millican, K.; Reynolds, M.; et al. 2022. Flamingo: a visual language model for few-shot learning. In *NeurIPS*.

Anderson, P.; He, X.; Buehler, C.; Teney, D.; Johnson, M.; Gould, S.; and Zhang, L. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*.

Andreas, J.; Rohrbach, M.; Darrell, T.; and Klein, D. 2016. Neural module networks. In *CVPR*.

Aubakirova, D.; Gerdes, K.; and Liu, L. 2023. PatFig: Generating Short and Long Captions for Patent Figures. In *ICCVW*.

Banerjee, S.; and Lavie, A. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *ACL Workshop*.

Bao, H.; Dong, L.; Piao, S.; and Wei, F. 2022. BEiT: BERT Pre-Training of Image Transformers. In *ICLR*.

Chen, C.; Zhang, R.; Kim, S.; Cohen, S.; Yu, T.; Rossi, R.; and Bunescu, R. 2019. Neural Caption Generation over Figures. In *UbiComp/ISWC*.

Chen, C.; Zhang, R.; Koh, E.; Kim, S.; Cohen, S.; and Rossi, R. 2020. Figure Captioning with Relation Maps for Reasoning. In *WACV*.

Chen, X.; Fang, H.; Lin, T.-Y.; Vedantam, R.; Gupta, S.; Dollár, P.; and Zitnick, C. L. 2015. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *ICLR*.

Esser, P.; Rombach, R.; and Ommer, B. 2021. Taming Transformers for High-Resolution Image Synthesis. In *CVPR*.

Gu, J.; Wang, J.; Cai, J.; and Jiang, H. 2019. Unified visual-semantic embeddings: Bridging vision and language with structured meaning representations. In *CVPR*.

Hochreiter, S.; and Schmidhuber, J. 1997. Long Short-Term Memory. *Neural Comput.*, 9(8): 1735–1780.

Hoffmann, J.; Borgeaud, S.; Mensch, A.; Buchatskaya, E.; Cai, T.; Rutherford, E.; de las Casas, D.; Hendricks, L. A.; Welbl, J.; Clark, A.; Hennigan, T.; Noland, E.; Millican, K.; van den Driessche, G.; Damoc, B.; Guy, A.; Osindero, S.; Simonyan, K.; Elsen, E.; Vinyals, O.; Rae, J. W.; and Sifre, L. 2022. An empirical analysis of compute-optimal large language model training. In *NeurIPS*.

Hsu, T.-Y.; Giles, C. L.; and Huang, T.-H. 2021. SciCap: Generating Captions for Scientific Figures. In *EMNLP 2021 (Findings)*.

Hu, E. J.; yelong shen; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *ICLR*.

Huang, Y.; Lv, T.; Cui, L.; Lu, Y.; and Wei, F. 2022. Layoutlmv3: Pre-training for document ai with unified text and image masking. In *ACM-MM*.

Kantharaj, S.; Leong, R. T.; Lin, X.; Masry, A.; Thakkar, M.; Hoque, E.; and Joty, S. 2022. Chart-to-Text: A Large-Scale Benchmark for Chart Summarization. In *ACL*.

Kay, A. 2007. Tesseract: an open-source optical character recognition engine. *Linux J.*, 2007(159): 2.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*.

Li, J.; Li, D.; Xiong, C.; and Hoi, S. C. H. 2022. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. In *ICML*.

Lin, C.-Y. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, 74–81. Barcelona, Spain: Association for Computational Linguistics.

Lin, T.-Y.; Maire, M.; Belongie, S. J.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft COCO: Common Objects in Context. In *ECCV*.

Liu, H.; Li, C.; Li, Y.; and Lee, Y. J. 2024a. Improved baselines with visual instruction tuning. In *CVPR*.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023. Visual Instruction Tuning. *ArXiv*, abs/2304.08485.

Liu, Y.; Yang, B.; Liu, Q.; Li, Z.; Ma, Z.; Zhang, S.; and Bai, X. 2024b. TextMonkey: An OCR-Free Large Multimodal Model for Understanding Document. *arXiv preprint arXiv:2403.04473*.

Liu, Z.; Hu, X.; Zhou, D.; Li, L.; Zhang, X.; and Xiang, Y. 2022. Code Generation From Flowcharts with Texts: A Benchmark Dataset and An Approach. In *EMNLP (Findings)*, 6069–6077.

Lu, J.; Xiong, C.; Parikh, D.; and Socher, R. 2017. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *CVPR*.

Lu, J.; Yang, J.; Batra, D.; and Parikh, D. 2018. Neural baby talk. In *CVPR*.

Luo, C.; Shen, Y.; Zhu, Z.; Zheng, Q.; Yu, Z.; and Yao, C. 2024. LayoutLLM: Layout Instruction Tuning with Large Language Models for Document Understanding. In *CVPR*.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. BLEU: a method for automatic evaluation of machine translation. In *ACL*.

Ramesh, A.; Pavlov, M.; Goh, G.; Gray, S.; Voss, C.; Radford, A.; Chen, M.; and Sutskever, I. 2021. Zero-shot text-to-image generation. In *ICML*.

Ren, S.; He, K.; Girshick, R. B.; and Sun, J. 2015. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *NeurIPS*.

Rodriguez, J. A.; Vazquez, D.; Laradji, I.; Pedersoli, M.; and Rodriguez, P. 2023. OCR-vqgan: Taming text-within-image generation. In *WACV*.

Shetty, R.; Roumeliotis, G.; and Laaksonen, J. 2017. Speaking the same language: Matching machine to human captions for image captioning. In *ICCV*.

Shibata, Y.; Kida, T.; Fukamachi, S.; Takeda, M.; Shinohara, A.; and Shinohara, T. 1999. Byte Pair Encoding: A Text Compression Scheme That Accelerates Pattern Matching. *ResearchGate*.

Shukla, S.; Gatti, P.; Kumar, Y.; Yadav, V.; and Mishra, A. 2023. Towards Making Flowchart Images Machine Interpretable. In *ICDAR*.

Sidorov, O.; Hu, R.; Rohrbach, M.; and Singh, A. 2020. Textcaps: a dataset for image captioning with reading comprehension. In *ECCV*.

Suzgun, M.; Melas-Kyriazi, L.; Sarkar, S.; Kominers, S. D.; and Shieber, S. 2024. The harvard uspto patent dataset: A large-scale, well-structured, and multi-purpose corpus of patent applications. In *NeurIPS*.

Tang, B.; Boggust, A.; and Satyanarayan, A. 2023. VisText: A Benchmark for Semantically Rich Chart Captioning. In *ACL*.

Team, G.; Anil, R.; Borgeaud, S.; Wu, Y.; Alayrac, J.-B.; Yu, J.; Soricut, R.; Schalkwyk, J.; Dai, A. M.; Hauth, A.; et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; Rodriguez, A.; Joulin, A.; Grave, E.; and Lample, G. 2023. LLaMA: Open and Efficient Foundation Language Models. *ArXiv*, abs/2302.13971.

Vinyals, O.; Toshev, A.; Bengio, S.; and Erhan, D. 2015. Show and tell: A neural image caption generator. In *CVPR*.

Wang, J.; Yang, Z.; Hu, X.; Li, L.; Lin, K.; Gan, Z.; Liu, Z.; Liu, C.; and Wang, L. 2022a. GIT: A Generative Image-to-text Transformer for Vision and Language. *Transactions on Machine Learning Research*.

Wang, P.; Yang, A.; Men, R.; Lin, J.; Bai, S.; Li, Z.; Ma, J.; Zhou, C.; Zhou, J.; and Yang, H. 2022b. OFA: Unifying Architectures, Tasks, and Modalities Through a Simple Sequence-to-Sequence Learning Framework. In *ICML*.

Xu, Y.; Li, M.; Cui, L.; Huang, S.; Wei, F.; and Zhou, M. 2020. Layoutlm: Pre-training of text and layout for document image understanding. In *KDD*, 1192–1200.

Xu, Y.; Xu, Y.; Lv, T.; Cui, L.; Wei, F.; Wang, G.; Lu, Y.; Florencio, D.; Zhang, C.; Che, W.; et al. 2021. LayoutLMv2: Multi-modal Pre-training for Visually-rich Document Understanding. In *ACL-IJCNLP*.

Yang, Z.; Dabre, R.; Tanaka, H.; and Okazaki, N. 2023. SciCap+: A Knowledge Augmented Dataset to Study the Challenges of Scientific Figure Captioning. *ArXiv*, abs/2306.03491.

Ye, J.; Hu, A.; Xu, H.; Ye, Q.; Yan, M.; Xu, G.; Li, C.; Tian, J.; Qian, Q.; Zhang, J.; et al. 2023a. UReader: Universal OCR-free Visually-situated Language Understanding with Multimodal Large Language Model. In *EMNLP 2023 (Findings)*.

Ye, Q.; Xu, H.; Xu, G.; Ye, J.; Yan, M.; Zhou, Y.; Wang, J.; Hu, A.; Shi, P.; Shi, Y.; et al. 2023b. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*.

Ye, Q.; Xu, H.; Ye, J.; Yan, M.; Hu, A.; Liu, H.; Qian, Q.; Zhang, J.; and Huang, F. 2024. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. In *CVPR*.

Zhang, J.; Zhao, Y.; Saleh, M.; and Liu, P. 2020. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *ICML*.

Zhu, D.; Chen, J.; Shen, X.; Li, X.; and Elhoseiny, M. 2024. MiniGPT-4: Enhancing Vision-Language Understanding with Advanced Large Language Models. In *ICLR*.