AI-FIGURES: A Fine-grained Task-oriented Dataset for developing Multimodal Scientific Literature Understanding

Anonymous ACL submission

Abstract

Diagrams and figures are a powerful medium of communication in scientific research. There is a recent spark in interest in the development of Machine Learning-driven applications involving scientific figures such as multimodal question answering, multimodal document retrieval, text-to-image generation or image captioning. Challenging tasks in this domain may be dependent only on a specific category of scientific figures. But there are no datasets in prior literature which provide a domain-specific broad classification of scientific figures. To fill this gap, we introduce AI-FIGURES a large scale dataset containing scientific figure-caption pairs which are classified into 9 different categories. We create this dataset by leveraging the idea of image segmentation and classification using the YOLO model. Our automated data acquisition pipeline can be implemented on other datasets also in order to classify their figures. We benchmark 6 Large Language Vision models and 5 Large Language models on our dataset for various tasks such as figure captioning, tag classification, text-to-figure generation, multimodal question answering and multimodal document retrieval. We show that there is a significant increase in a model's inference capabilities when we finetune it on our dataset. Our dataset and code will be released in the final version.

1 Introduction

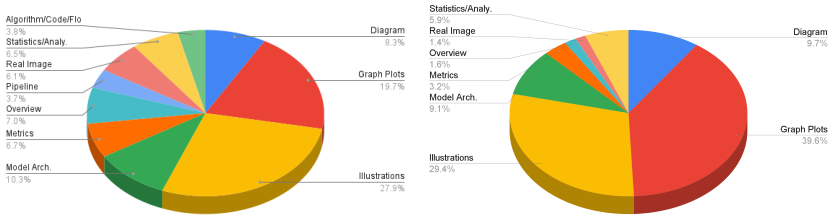
Images create a visual imprint on our brain that is immediately able to trigger the human perceptual system to process the simultaneous conceptual representation. Images serve as vital elements in conveying crucial aspects of scholarly content too, such as methodological explanations, experimental results, and comparative analyses. Scientific figures encompasses diverse visual elements, which may be categorized as diagrams employing shapes and lines, charts using axes, labels, and data points, or images depicting real-word scenes (Huang et al., 2024). Recognizing the intrinsic importance of figures and tables, recent research endeavors have underscored the necessity of developing robust systems capable of extracting and interpreting these visual elements.

Vast strides have been made in multimodal tasks in the open domain like text-to image generation (Xu et al., 2018; Ramesh et al., 2021; Saharia et al., 2022; Ramesh et al., 2022; Rombach et al., 2022; Esser et al., 2024), multimodal document retrieval, multimodal document summarization (Jangra et al., 2023) and multimodal question answering (Masry et al., 2022; Yue et al., 2024; Lu et al., 2024).

Scientific figures with additional cues like their types and captions can prove quite useful in each of these tasks. For example, figures in a Computer Science research paper might be related to communication networks, computer architecture, graphs or line plots. For text-to-image synthesis, (Rodriguez et al., 2023b) filter figures by searching for keywords, such as "architecture", "model diagram" or "pipeline," in their captions. Clearly, it would be useful if there is a large corpus of scientific figures with a fine-grained classification. It would be a useful resource in other multimodal tasks as well.

To address this need, in this paper we introduce AI-FIGURES (Figure 1), which is large fine grained dataset obtained through a YOLO-based distantly supervised pipeline. Our dataset has 9 different categories for representing various kinds of scientific figures that are particularly common in Artificial Intelligence research papers.

We evaluate a wide spectrum of pre-trained foundational models on our proposed dataset for a diversity of vision-to-text and text-to-vision tasks. Our experiments demonstrate the challenging nature as well as the effectiveness of our dataset. The challenging nature is exhibited by the low results obtained by state-of-the-art Large Vision Language Models (LVLMs) on the standard "figure captioning" and the relatively new "text-to-figure" tasks.



(a) AI-FIGURES-HUMAN (b) AI-FIGURES

Figure 1: The class-wise distribution in the human annotated and the distantly-supervised AI-FIGURES dataset.

We show the effectiveness of our dataset as a training resource which can improve the scientific literature understanding capability of LVLMs.

Our contributions are the following:
(a) We introduce AI-FIGURES, a multimodal dataset that is expected to aid researchers in modeling new tasks in scientific literature understanding that are dependent on figure types. We also present AI-FIGURES-HUMAN, a corpus of scientific figures and captions, that is manually annotated and has been used to distantly supervise the annotation of AI-FIGURES.
(b) We analyze three different tasks using our dataset which demonstrate the limitations of pretrained LVLMs in scientific literature understanding.
(c) We demonstrate that training on our dataset can lead to performance improvement on tasks such as multimodal question-answering and multimodal document retrieval.

2 AI-FIGURES-HUMAN

We introduce a human-annotated dataset comprising of a corpus of figures in the Artificial Intelligence/Machine Learning domain paired with their textual contexts, i.e., figure captions. We have labeled bounding boxes for figure and caption regions for each document page. The Roboflow Annotate[1] platform was used to assist annotators to mark the bounding regions. This platform facilitated dataset pre-processing, division into train, validation, and test sets. Human annotations were performed on a set of 200 research documents, with 100 each from ACL Anthology (Annual Meeting of the Association for Computational Linguistics)[2] and CVPR (IEEE/CVF Conference on Computer Vision and Pattern Recognition)[3]. The final dataset stood at 4975 images split into training (3790 images), validation (803 images) and test (382 images) sets.

We have designed our schema keeping in mind the taxonomy of figures of research papers in the Computer Science domain. The 10 figure category classes and 1 caption class that were curated for the inferential segregation of the figures are:

The **Algorithm/Code/Flowchart** class contains figures involving flowcharts, code snippets and pseudo-code algorithm outlines. The **Diagram** category consists of abstract schematic representations with labeling. The **Graph plots** class shows non-performance and non-statistical plotting. The **Illustrations and Examples** category represents the visual depiction of an idea or feature. The **Model architecture** class, in the context of Artificial Intelligence, shows a detailed probe into a machine learning model structure. The **Model Performance with Metrics** class represents plots of baselines, plots of variations of different metrics with training. The **Overview/Procedure** class comprises figures showing high-level glances at the structural and functional aspects of the proposed technique or the step-wise details of a procedure. The **Pipeline** class contains figures representing a step by step workflow showing the organization or the ideation of a topic. The **Real Image** category comprises real-world images which may be either instances from a dataset used in the research paper or any other image in the open-domain. The **Statistics and Analysis** category Distributions of parameters, Statistical variations, Ablation study results and Analytical experimentation. **Captions** are also segmented and put into a common class for all figure captions.

The annotation guidelines that have been used for the human annotation of **AI-Figures** are provided in Appendix B.

[1] https://roboflow.com/annotate
[2] https://aclanthology.org/
[3] https://openaccess.thecvf.com/

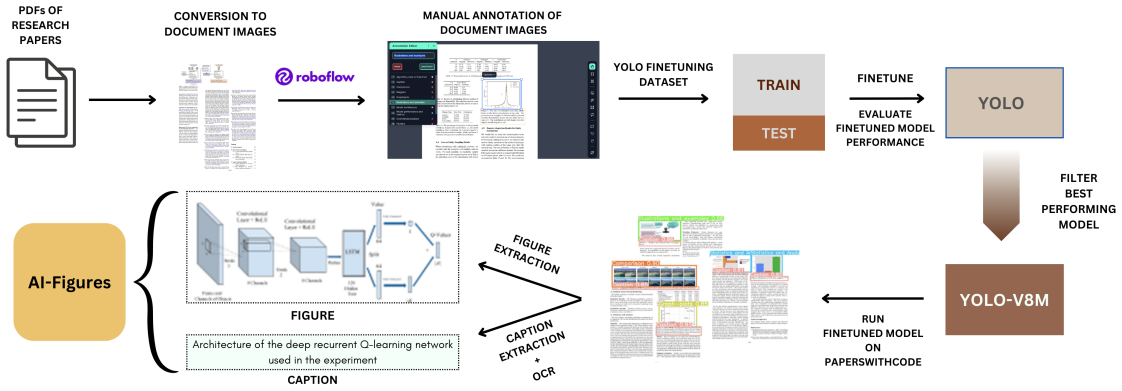


Figure 2: The construction pipeline used for the AI-FIGURES dataset.

3 AI-FIGURES

We now introduce the process of large-scale extraction of figures and captions from scientific papers in a distantly supervised fashion based on AI-FIGURES-HUMAN. Figure 2 shows the entire construction pipeline to create AI-FIGURES.

3.1 Dataset Construction Methodology

3.1.1 PDF to Images

Our dataset construction process leverages the idea that if we consider a single page of a research paper we only need to segment the area containing the figure and the small chunk of text that is most adjacent to it. Therefore, we handle the figure and caption extraction task with an object detection pipeline. Each page of every PDF document is converted into an image using a Python program that utilized the *pdf2image*[4] library. Thus, the figures and captions in the page are only objects in the image. This would allow us to assign separate classes to each figure. Captions, which are present alongside figures, are naturally treated as objects as well and can be classified into a separate class.

3.1.2 Object Detection with YOLO

YOLO (You Only Look Once) (Redmon et al., 2015) is a hugely popular fast object detection and image segmentation model, that was initially released in 2015. In YOLO, object detection is reformulated as a regression problem from image pixels to bounding box coordinates and class probabilities. The original YOLO model consisted of a single convolutional network which simultaneously predicts multiple bounding boxes and class probabilities for those boxes on full images.

[4] https://pypi.org/project/pdf2image/

Model	mAP	mAP 0.5-0.95	P	R
YOLOv5s	0.506	0.434	0.461	0.607
YOLOv5m	0.497	0.449	**0.471**	0.558
YOLOv5l	0.51	0.458	0.434	0.644
YOLOv8s	0.49	0.441	0.424	0.62
YOLOv8m	**0.515**	**0.462**	0.445	0.667
YOLOv8l	0.505	0.456	0.42	**0.695**

Table 1: Results of YOLO on AI-FIGURES (Human). P represents Precision while R represents Recall

We train several versions of YOLO models including YOLOv5s, YOLOv5m, YOLOv5l, YOLOv8s, YOLOv8m and YOLOv8l on AI-FIGURES-HUMAN. Based on the mean Average Precision (mAP) scores on the test set of AI-FIGURES-HUMAN, as shown in Table 1, we select YOLOv8m for figure and caption extraction on the larger corpus.

3.2 Data Collection

We use the URLs present in the *PapersWithCode*[5] repository to curate a corpus of open-access research papers as PDF documents. The open-sourced corpus covers a wide variety of research papers in the AI-ML domain across multiple conferences and journal. We use the YOLOv8m model to extract figures from them. Subsequently, we run an OCR (Optical character recognition) model[6] over the Captions objects to convert them to texts.

3.3 Dataset Refinement Process

After manual assessment of the extracted figures and captions, two issues were revealed with our above approach. Firstly, if the image of the page

[5] https://paperswithcode.com/
[6] https://github.com/tesseract-ocr/tesseract

from a document contains two or more figures belonging to the same category, the YOLO model extracts only the last extracted figure despite the fact that it detects both the figures. However, the model extracts all the captions in the input page image. This leads to a mismatch in the number of captions and images. To circumvent this problem, we map the selected figure bounding box co-ordinate to the bounding box co-ordinate of the closest caption by calculating the Euclidean distance between the centres of the bounding boxes.

Secondly, if a detected figure crop has been classified into multiple classes at the same time with varying confidence scores, then the YOLO model allots the figure to both classes. To remove such ambiguity, we first detect multiple class assignments based on the maximal overlap of bounding box co-ordinates. We then assign the class with the greatest confidence score to the figure.

Dataset Cleaning: We remove all figures which have captions shorter than 5 words. Also, phrases like Figure x:/Figure x./Fig. x:/ Fig. x are deleted from the beginning of each caption.

Finally, we remove the **Algorithm/Code/Flowchart** class from the dataset due to the high occurrence of hallucinations in this category. The frequent hallucinations arise because the model often confuses an Algorithm/Code image with a regular text snippet.

3.4 Dataset Statistics

Our final dataset contains $1,33,749$ scientific figure-caption pairs. We present the class-wise statistics of both the human-annotated dataset and the larger inferred dataset in Table 2. Figure 3 shows the distribution of document sources in AI-FIGURES. Our dataset contains a total of 4,925,626 words with the average caption length being 36.83 words and the quartile length being $[13, 27, 49]$.

3.5 Construction Approach Comparison

PDFFigures: The original approach (Clark and Divvala, 2015) is based on the analysis of documents pages and has three phases: Caption Detection using keyword search, Region Identification using paragraph grouping with classification and Figure Assignment using a scoring function to rate the proposed regions. We use the PDFFigures 2.0 version (Clark and Divvala, 2016) for the purpose of testing which extends the original algorithm for a wider variety of paper formats.

Model	AI-FIGURES-HUMAN	AI-FIGURES
Algo./Flowchart	183	-
Diagram	402	12,975
Graph Plots	956	52,932
Illustrations	1,351	39,359
Model Arch.	500	12,169
Metrics	324	4,305
Overview	340	2,095
Pipeline	179	59
Real Image	296	1,910
Stat./Analysis	313	7,945
Total	**4,844**	**133,749**

Table 2: Class-wise statistics of AI-FIGURES-HUMAN and AI-FIGURES

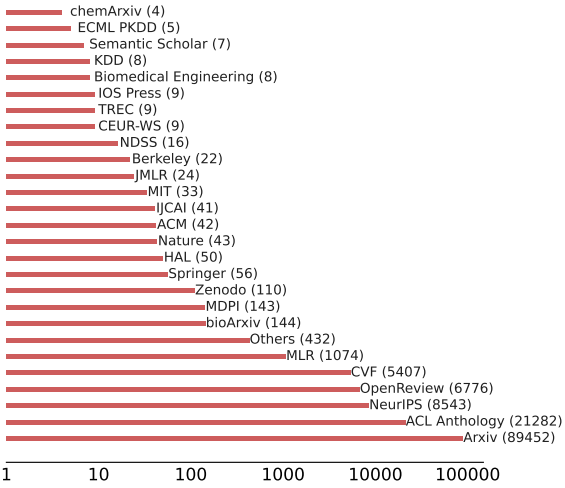


Figure 3: Domain Distribution of the figures in our dataset

We test the approach of PDFFigures 2.0 with our construction pipeline on the test set of our human-annotated dataset, AI-FIGURES-HUMAN. The results are present in Table 3, where we see that our method comprehensively outperforms the PDFFigures approach on all metrics. Upon qualitative evaluation, we find that there are two major reasons for the performance of PDFFigures, firstly there are a lot of tables extracted along with the figures and secondly, this algorithm randomly extracts many blank strips.

PaperMage (Lo et al., 2023): It is an open-source Python toolkit which allows the representation and manipulation of both textual and visual elements in a document.

In the test set used for evaluating PaperMage there were 5532 PDF page images out of which 4325 pages contained figures. In 55 out of 4325 pages, PaperMage showed some signs of figure recognition. In the remaining 4270 pages, no figures were detected, indicating false negatives

Model	P	R	F1	Avg. P (IOU 0.5)
PDFFigures	0.395	0.634	0.487	0.333
YOLOv8m	**0.445**	**0.667**	**0.534**	**0.515**

Table 3: P represents Precision while R represents Recall

across these pages. In 27% of the 55 pages, PaperMage exhibits poor extraction quality, with the bounding box placed in the middle of the figure, failing to properly define the figure's boundaries. As a result, most of these extractions are sliced and unsuitable for evaluation. However, in 41 images, the extractions were decent and suitable for evaluation, where we observed an average Intersection over Union (IoU) score of 0.6818.

3.6 Human Evaluation

We construct the following manual evaluation setup to evaluate our dataset construction process. We randomly construct 6 different sets with 100 figure-caption-category triplets in each of them. Each annotator is provided with the extracted figure, the extracted caption, the selected category and also the URL to the original paper PDF from where they have been extracted.

We select 6 graduate students with knowledge in Computer Science as annotators to evaluate our distantly-supervised dataset. We ask the annotators to categorize the dataset samples into the following four categories: (1) **Acceptable**, where the image segmentation is done correctly, the figure is categorized into an acceptable class and the caption is extracted correctly; (2) **Figure Segmentation Error**, where the figure crop is done incorrectly; (3) **Figure Classification Error**, where the model inaccurately classifies the figure into an unrelated category; (4) **Figure-Caption Pairing Error**, where the figure is paired with an incorrect caption.

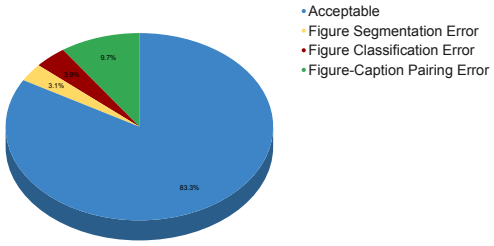


Figure 4: Results for the manual analysis of our distantly supervised dataset, AI-FIGURES

Figure 4 shows the aggregated results of the manual evaluation of the dataset construction pipeline by human annotators. We see that in most cases the dataset samples are in the *Acceptable* category. The Figure-Caption pairing error is the largest contributor to the error list, followed by the classification and segmentation errors, respectively.

4 Comparison with Related Datasets

Table 4 contains a list of related datasets and shows them in comparison with our dataset. CS-150 (Clark and Divvala, 2015) a human-annotated dataset containing 150 Computer Science papers with the ground-truth labels demarcating the locations of the figures, tables and captions within them. The CS-Large dataset (Clark and Divvala, 2016) comprises annotations from 346 papers. The Paper2Fig100k (Rodriguez et al., 2023b) dataset contains 102, 453 images from 183, 427 papers that were downloaded from arXiv in areas of Machine Learning, Artificial Intelligence, Computer Vision and Pattern Recognition, and Computation and Language. The images were extracted from the documents using the popular GROBID tool. Figures in Paper2Fig100k are not labeled into named classes. The Multimodal ArXiv dataset (Li et al., 2024b) has also been extracted from ArXiv but on a larger domain set including 32 domains. This dataset contains a subset called ArXivCap which consists 6.4M images and approximately 3.9M main captions. VisImages (Deng et al., 2022) presents 12,267 images with captions from IEEE conferences, but they are limited to graph plots. ACL-FIG (Karishma et al., 2023) contains 112,05 unlabeled figures from 55,760 papers in ACL Anthology. It is accompanied with its labeled subset ACL-FIG-PILOT, that contains 1671 scientific figures with 19 manually verified labels. However, ACL-FIG figures do not contain captions, and this limits their utility.

5 Downstream tasks

5.1 Figure Captioning

Single figure captioning for scientific figures aims to capture the complex architectures, illustrations and data trends in a concise yet informative manner. Therefore, given a figure F and an instruction prompt P, a chosen model $\mathcal{M}$ is required to generate a suitable caption $\widehat{C}$ for F:

$$\widehat{C} = \mathcal{M}(F, P) \quad (1)$$

Dataset	Source	Annotation	Papers	Figures	Classes
CS-150 (Clark and Divvala, 2015)	CS-conferences	Manual	150	458	✗
CS-Large (Clark and Divvala, 2016)	Semantic Scholar	Manual	346	952	✗
Paper2Fig100k (Rodriguez et al., 2023b)	ArXiv	GROBID	183,427	102,453	✗
ArXivCap (Li et al., 2024b)	ArXiv	ImageMagick	572K	6.4M	✗
ACL-FIG-PILOT (Karishma et al., 2023)	ACL Anthology	Manual	-	1,671	19
ACL-FIG (Karishma et al., 2023)	ACL Anthology	-	55,760	112,052	✗
VisImages (Deng et al., 2022)	IEEE InfoVis and VAST	Manual	1,397	12,267	34
AI-FIGURES-HUMAN	PaperswithCode	Manual	200	4,844	10
AI-FIGURES	PaperswithCode	YOLO	26,969	133,749	9

Table 4: Comparison with prior scientific figure datasets.

Model	Zero-shot Captioning			Context = Title			Context = Title + Abstract		
	BLEU-2	R-L	B-S	BLEU-2	R-L	B-S	BLEU-2	R-L	B-S
MOLMO-7B	1.42	8.13	81.41	1.27	7.99	81.49	1.35	7.91	81.61
InternVL2_5-8B	1.31	7.83	81.01	1.40	7.96	81.18	1.15	7.09	80.83
Qwen2-VL-7B	1.91	9.00	81.40	**2.35**	**9.62**	81.92	**2.28**	**9.50**	**81.75**
MiniCPM-V	**1.94**	**9.54**	**81.66**	1.47	8.53	**82.38**	1.64	7.56	81.07
Janus-Pro-7B	1.60	8.59	81.10	1.62	8.73	81.22	1.79	8.86	81.42

Table 5: Evaluation results of the Figure Captioning task. R-L refers to the ROUGE-L score and B-S refers to BERTscore

Model	BLEU-2	Rouge-L	BERTscore
GIT-base	1.58	**10.74**	**83.22**
GIT-large	**3.01**	10.01	81.61

Table 6: Results for finetuning for captioning

$\widehat{C}$ is then compared to the original caption C and its quality is assessed. To provide more context to the model, we propose a modified version of this task, where we provide the model $\mathcal{M}$ with metadata from the research paper such as the title t and the abstract a. This tests whether additional in-domain information relating to the figure F can aid in the task.

We benchmark the following Large Vision Language models on the figure-captioning task: MOLMO-7B (Deitke et al., 2024), InternVL2_5-8B (Chen et al., 2024), Qwen2-VL-7B (Wang et al., 2024), Janus-Pro-7B (Chen et al., 2025) and MiniCPM-V (Yao et al., 2024). For each model, we report the BLEU-2 (Papineni et al., 2002), ROUGE-L (Lin, 2004) and the BERT-Score (Zhang et al., 2019) in all the three settings, i.e., captioning without context, captioning with title as context, and captioning with both title and abstract as context. We also fine-tune the GIT-base and GIT-large (Wang et al., 2022) models on the uncleaned version of our dataset so that me may train on as many figure caption pairs as possible, but while testing we use the cleaned version.

Table 5 and Table 6 show the results of the various models on the figure captioning task. We see that in spite of being very proficient in the captioning task in the open-domain, the LVLMs perform poor on scientific figures, which shows that there is lot of scope for improvement on this task. Fine-tuning the GIT models provide slightly better results than fine-tuning the LVLMs.

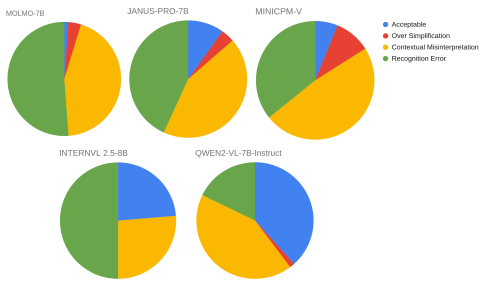


Figure 5: Manual Analysis for Figure Captioning

Manual Analysis: We also construct a manual evaluation setup for the Figure Captioning task. We construct two different sets with 25 captions each and ask three annotators to analyze the quality of the generated captions for each model. Each annotator is provided with the figure, the gold standard caption from our dataset and the generated captions from the MOLMO-7B (Deitke et al., 2024), InternVL2_5-8B (Chen et al., 2024), Qwen2-VL-7B (Wang et al., 2024), Janus-Pro-7B (Chen et al., 2025) and the MiniCPM-V (Yao et al., 2024) models. We select two doctoral students who work in allied areas and have at least one publication in the domain of Artificial Intelligence/Machine Learning/Computer Vision/Natural Language Processing as the annotators for this task. The details about the

Model	Training Corpus	FID ↓	IS ↑	KID ↓	LongCLIP image-image sim.
SDXL_base_1.0	Zero-shot inference	96.15	**9.655± 0.641**	0.069± 0.002	0.64
SDXL_base_1.0	AI-FIGURES	**84.38**	6.406 ± 0.314	**0.062± 0.002**	**0.67**

Table 7: Text-to-Figure

Model	All	Modality		Reasoning Type			
		Table	Figure	COM	DE	LOC	VU
GPT-4o	**0.5**	**0.52**	**0.454**	**0.443**	**0.565**	**0.57**	**0.418**
Qwen2-VL-7B-Instruct	0.1334	0.112	0.1863	0.1233	0.1135	0.1667	0.1708
Qwen2-VL-7B-Instruct(fine-tuned)	0.2021	0.1848	0.2446	0.2187	0.1727	0.2272	0.1966

Table 8: Mean reciprocal rank (MRR) on M3SciQA. Reasoning types: **COM**: comparison, **DE**: data extraction, **LOC**: location, **VU**: visual understanding. GPT-4o results are taken from the original paper. The second best results are underlined.

annotation and the annotation guidelines for this task are provided in Appendix D.

In line with the quantitative results, we see in Figure 5 that the there are only a few acceptable responses across all models with the Qwen model performing the best among the given models, whereas MOLMO performs the worst.

5.2 Text-to-Figure

Based on the success of text-to-image (T2I) generation, there have have been some introductory trials in the scientific and allied domains to generate figures and diagrams (Zala et al., 2023). In the text-to-figure task, when a generator model $\mathcal{M}$ is presented the image caption C as a textual prompt, it is required to generate the corresponding figure $\widehat{F}$, which is then compared to the original figure F,

$$\widehat{F} = \mathcal{M}(C) \tag{2}$$

We select the Diagram, Model Architecture, Overview/Procedure and the Pipeline categories from our dataset to create a training set comprising of $21,839$ images and the test set with $5,461$ images. We then fine-tune the Stable Diffusion-XL model for 20 epochs with a batch size of 8 and a learning rate of 1e-06. We compute the Fréchet Inception Distance (FID) (Heusel et al., 2017), Inception Score (Salimans et al., 2016), the Kernel Inception Distance (Bińkowski et al., 2021) and the LongCLIP image-image similarity metrics (Regenwetter et al., 2023) for this task.

Table 7 shows the results for the text-to-figure task. Quantitatively, the results are better than that obtained by (Rodriguez et al., 2023a). However, manual review shows that the generated images are hardly comprehensible. A set of images that are received as output from the zero-shot and the fine-tuned models are presented in the Appendix.

5.3 Tag Classification

We introduce a task in which we test the capability of a pre-trained language model to deduce the type of the figure when it is provided with only the caption associated with the figure.

We test it on our dataset by inferencing on LLMs including Llama-3.2-1B, Llama-3.1-8B (Grattafiori et al., 2024), Mistral-7B (Jiang et al., 2023), Qwen2.5-0.5B and Qwen2.5-7B (Team, 2024).

$$\widehat{T} = \mathcal{M}(C) \tag{3}$$

Model	Prec.	Recall	F1
Llama-3.2-1B-Instruct	24.10	12.71	12.94
Llama-3.1-8B-Instruct	42.66	20.82	20.84
Mistral-7B-Instruct-v0.2	49.81	26.84	24.66
Qwen2.5-0.5B-Instruct	37.14	**39.27**	31.50
Qwen2.5-7B-Instruct	**54.51**	35.76	**39.33**

Table 9: Tag Classification

The results for the tag-classification task are present in Table 9, wherein we see that Qwen family of LLMs perform best followed by the Mistral and the LLaMa models. We also see that this task proves to be challenging for the LLMs since no model has even crossed the 40 F1 score mark.

6 Improving LVLMs with AI-FIGURES

Finetuning Hypothesis: We choose a certain subset of categories from our dataset and then fine-tune a LVLM on this subset. We posit that the finetuned model will work better than the original model.

Experimental Setup We use the "Graph Plots" and the "Statistics and Analysis" classes to form a

Data	Qwen2-VL-7B	QWEN2-VL-7B (Finetuned)
PaperQA	53.85	**61.54**
ScienceQA	55.56	55.56
IQTest	25	**50**
TabMWP	44.44	**55.56**
ChartQA	66.67	**83.33**

Table 10: Accuracy score over Multiple Choice questions (MCQs) in the MathVista dataset

Data	Qwen2-VL-7B	QWEN2-VL-7B (Finetuned)
CLEVR-Math	67.74	**72.58**
DVQA	82.26	**85.48**
TabMWP	26.42	**43.4**

Table 11: Accuracy score over Freeform questions in the MathVista dataset

subset of our entire dataset i.e. we choose a combined $48,716$ figures from our training set to create this subset. We then prompt the InternVL2_5-8B (Chen et al., 2024) model to generate 3 unique sets of question-answer pairs for each figure. We then finetune the Qwen2-VL-7B (Wang et al., 2024) on this derived question answering set. We use QLoRA in the HuggingFace Ecosystem (TRL) to train the model for 10 epochs with a train batch size of 4, learning rate of 2e-04, maximum sequence length of 1024. MathVista (Lu et al., 2024) is a mathematical reasoning benchmark within visual contexts. We show the performance of five subsets of MCQ questions and three subsets of Free form questions of the MathVista dataset which are present in its testmini version. We choose the subsets such that they align with our problem setup i.e. they are dependent on either academic papers or charts or scientific knowledge and require numerical rationale.

Tables 10 and 11 show the results on the five MCQ subsets and three free form subsets of the MathVista dataset. Clearly, domain-specific finetuning helps in achieving better results.

6.1 Multimodal Document Retrieval

This task necessitates both multimodal and multi-document reasoning over scientific papers. M3SCIQA (Li et al., 2024a) is a benchmark which contains expert-annotated questions from paper clusters. The questions are divided into four reasoning categories: comparisons, data extraction, locations and visual understanding. Therefore, given a locality-specific question Q, the corresponding image I and the list of documents $D = \{d_1, d_2, ..., d_n\}$, the task is to determine the ranking $R = \{r_1, r_2, ..., r_n\}$ of papers based on the relevance of D to Q and I.

$$R = \mathcal{M}(Q, I, D) \tag{4}$$

We only consider the locality-specific document retrieval setup here, since it tests the capability of VLVMs. Table 8 shows the results for this task. The fine-tuned model outperforms the naive model, supporting our hypothesis.

7 Related work

There have been some introductory work in the area of scientific figure and caption extraction. Most of them include the extraction of tables within their ambit, and consider tables as a form of figures. Almost none of these methods propose an taxonomy for the categorization of the extracted figures.

Figure extraction: Software tools which are ideal for the off-the-shelf-processing of scientific documents include GROBID (GRO, 2008–2024), ParsCit (Isaac Councill and Kan, 2008) and CERMINE (Tkaczyk et al., 2015). They use varios Machine Learning algorithms like CRFs (Conditional Random Fields), recurrent neural networks, and even recent deep learning models. PDFFigures (Clark and Divvala, 2015), a widely used figure extractor, performs structural analysis of individual pages of a document and can identify, with high accuracy, figures, tables, and captions in the pages.

Related datasets: Datasets of figure-caption pairs in the domain of scientific literature typically focus only on scientific plots. Example datasets include FigureQA (Kahou et al., 2017), DVQA-cap (Kafle et al., 2018), FigJAM (Qian et al., 2021), SciCap (Hsu et al., 2021), Paper2Fig100k (Rodriguez et al., 2023b), and ACL-FIG (Karishma et al., 2023). While the first 4 of these datasets exclusively contain graph plots like line plots, barcharts, etc., the remaining has more diverse figures.

8 Conclusion

We introduce the AI-FIGURES dataset in this paper. We also propose a construction pipeline which can be used to extract and label figure-caption pairs. Our dataset is divided into fine-grained categories, which makes it possible to use it on category-specific tasks. We show the challenging nature of the captioning, text-to-figure and the tag classification tasks. We also show the improvements achieved on fine-tuning a LVLM on our dataset.

Limitations

We hereby state the limitations of our work. We understand that the scientific domain is extremely challenging and large, and Artificial Intelligence, i.e., the area we choose for the creating the dataset here is a very niche and evolving area. So the dataset may need to be regularly updated for high practical utility to researchers. Since we use distant supervision, the AI-FIGURES dataset is likely to contain some errors. Nevertheless, we believe that our dataset construction pipeline can be used in any domain very easily.

Furthermore, the space of language-vision models and language models is rapidly evolving and therefore, we have not been able to exhaustively test on many of these models.

References

2008–2024. Grobid. https://github.com/kermitt2/grobid.

Mikołaj Bińkowski, Danica J. Sutherland, Michael Arbel, and Arthur Gretton. 2021. Demystifying mmd gans. *Preprint*, arXiv:1801.01401.

Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. 2025. Janus-pro: Unified multimodal understanding and generation with data and model scaling. *arXiv preprint arXiv:2501.17811*.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. 2024. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *Preprint*, arXiv:2312.14238.

Christopher Clark and Santosh Divvala. 2016. Pdffigures 2.0: Mining figures from research papers. In *Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries*, JCDL '16, page 143–152, New York, NY, USA. Association for Computing Machinery.

Christopher Clark and Santosh Kumar Divvala. 2015. Looking beyond text: Extracting figures, tables and captions from computer science papers. In *AAAI Workshop: Scholarly Big Data*.

Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, et al. 2024. Molmo and pixmo: Open weights and open data for state-of-the-art vision-language models. *arXiv preprint arXiv:2409.17146v2*. Allen Institute for AI and University of Washington.

Dazhen Deng, Yihong Wu, Xinhuan Shu, Jiang Wu, Siwei Fu, Weiwei Cui, and Yingcai Wu. 2022. Visimages: A fine-grained expert-annotated visualization dataset. *IEEE Transactions on Visualization and Computer Graphics*, 29(7):3298–3311.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, and Robin Rombach. 2024. Scaling rectified flow transformers for high-resolution image synthesis. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 12606–12633. PMLR.

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

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Ting-Yao Hsu, C Lee Giles, and Ting-Hao Huang. 2021. SciCap: Generating captions for scientific figures. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3258–3264, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jiani Huang, Haihua Chen, Fengchang Yu, and Wei Lu. 2024. From detection to application: Recent advances in understanding scientific tables and figures. *ACM Comput. Surv.*, 56(10).

C Lee Giles Isaac Councill and Min-Yen Kan. 2008. Parscit: an open-source crf reference string parsing package. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA). Http://www.lrec-conf.org/proceedings/lrec2008/.

Anubhav Jangra, Sourajit Mukherjee, Adam Jatowt, Sriparna Saha, and Mohammad Hasanuzzaman. 2023. A survey on multi-modal summarization. *ACM Comput. Surv.*, 55(13s).

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Kushal Kafle, Brian Price, Scott Cohen, and Christopher Kanan. 2018. DFQA: Understanding data visualizations via question answering. In *Proceedings of*

the IEEE conference on computer vision and pattern recognition, pages 5648–5656.

Samira Ebrahimi Kahou, Vincent Michalski, Adam Atkinson, Ákos Kádár, Adam Trischler, and Yoshua Bengio. 2017. FigureQA: An annotated figure dataset for visual reasoning. *arXiv preprint arXiv:1710.07300*.

Zeba Karishma, Shaurya Rohatgi, Kavya Shrinivas Puranik, Jian Wu, and C. Lee Giles. 2023. Acl-fig: A dataset for scientific figure classification. *Preprint*, arXiv:2301.12293.

Chuhan Li, Ziyao Shangguan, Yilun Zhao, Deyuan Li, Yixin Liu, and Arman Cohan. 2024a. M3SciQA: A multi-modal multi-document scientific QA benchmark for evaluating foundation models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15419–15446, Miami, Florida, USA. Association for Computational Linguistics.

Lei Li, Yuqi Wang, Runxin Xu, Peiyi Wang, Xiachong Feng, Lingpeng Kong, and Qi Liu. 2024b. Multimodal ArXiv: A dataset for improving scientific comprehension of large vision-language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14369–14387, Bangkok, Thailand. Association for Computational Linguistics.

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.

Kyle Lo, Zejiang Shen, Benjamin Newman, Joseph Chang, Russell Authur, Erin Bransom, Stefan Candra, Yoganand Chandrasekhar, Regan Huff, Bailey Kuehl, Amanpreet Singh, Chris Wilhelm, Angele Zamarron, Marti A. Hearst, Daniel Weld, Doug Downey, and Luca Soldaini. 2023. PaperMage: A unified toolkit for processing, representing, and manipulating visually-rich scientific documents. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 495–507, Singapore. Association for Computational Linguistics.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2024. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *Preprint*, arXiv:2310.02255.

Ahmed Masry, Xuan Long Do, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. 2022. ChartQA: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2263–2279, Dublin, Ireland. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 311–318, USA. Association for Computational Linguistics.

Xin Qian, Eunyee Koh, Fan Du, Sungchul Kim, Joel Chan, Ryan A. Rossi, Sana Malik, and Tak Yeon Lee. 2021. Generating accurate caption units for figure captioning. In *Proceedings of the Web Conference 2021*, WWW '21, page 2792–2804, New York, NY, USA. Association for Computing Machinery.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-shot text-to-image generation. In *International conference on machine learning*, pages 8821–8831. Pmlr.

Joseph Redmon, Santosh Kumar Divvala, Ross B. Girshick, and Ali Farhadi. 2015. You only look once: Unified, real-time object detection. *CoRR*, abs/1506.02640.

Lyle Regenwetter, Akash Srivastava, Dan Gutfreund, and Faez Ahmed. 2023. Beyond statistical similarity: Rethinking metrics for deep generative models in engineering design. *Computer-Aided Design*, 165:103609.

Juan A Rodriguez, David Vazquez, Issam Laradji, Marco Pedersoli, and Pau Rodriguez. 2023a. FigGen: Text to scientific figure generation. *arXiv preprint arXiv:2306.00800*.

Juan A Rodriguez, David Vazquez, Issam Laradji, Marco Pedersoli, and Pau Rodriguez. 2023b. OCR-VQGAN: Taming text-within-image generation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3689–3698.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, Xi Chen, and Xi Chen. 2016. Improved techniques for training gans. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

D. Tkaczyk, P. Szostek, M. Fedoryszak, P.J. Dendek, and Ł. Bolikowski. 2015. Cermine: automatic extraction of structured metadata from scientific literature. In *International Journal on Document Analysis and Recognition (IJDAR)*, pages 1433–2825.

Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. 2022. Git: A generative image-to-text transformer for vision and language. *Preprint*, arXiv:2205.14100.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. 2018. AttnGAN: Fine-grained text to image generation with attentional generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1316–1324.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, Qianyu Chen, Huarong Zhou, Zhensheng Zou, Haoye Zhang, Shengding Hu, Zhi Zheng, Jie Zhou, Jie Cai, Xu Han, Guoyang Zeng, Dahai Li, Zhiyuan Liu, and Maosong Sun. 2024. Minicpm-v: A gpt-4v level mllm on your phone. *Preprint*, arXiv:2408.01800.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of CVPR*.

Abhay Zala, Han Lin, Jaemin Cho, and Mohit Bansal. 2023. DiagrammerGPT: Generating open-domain, open-platform diagrams via llm planning. *arXiv preprint arXiv:2310.12128*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with BERT. *CoRR*, abs/1904.09675.

A Dataset Statistics

Our dataset contains the following fields:

- Figure Filename
- Figure Caption
- Paper Abstract
- Paper Title
- PDF URL

B Dataset Construction - Annotation Guidelines

- Each figure and its corresponding caption must have a separate bounding box.
  - Figures should be assigned to exactly 1 of the 10 predefined classes .
  - Captions are always assigned the class "Caption".
  - Ensure no overlap between figure and caption annotations.
- Bounding Box Rules
  - Draw tight bounding boxes around each figure and its caption.
  - The caption box should cover only the text of the caption, not surrounding text.
  - The figure box should include only the visual content of the figure, avoiding page borders or surrounding text.
- Classifying Figures
  - Carefully examine the content and concept behind each figure.
  - Assign the most appropriate class from the predefined categories listed below.
  - If a figure could belong to multiple categories, choose the most dominant or relevant one.
- Subfigures and Complex Figures
  - If a figure consists of multiple subfigures labeled as (a), (b), (c), etc., annotate the entire figure as one bounding box.
  - If subfigures have separate captions, annotate them individually with their respective captions.
- For the Algorithms Code or Flowchart class, ensure that only the code or flowchart is included in the bounding box, excluding body text explanations.

Model	AI-FIGURES-HUMAN	AI-FIGURES-Before Clean	AI-FIGURES
Algo./Flowchart	183	10,014	-
Diagram	402	14,704	12,975
Graph Plots	956	55,676	52,932
Illustrations	1,351	42,066	39,359
Model Arch.	500	15,839	12,169
Metrics	324	4,548	4,305
Overview	340	2,201	2,095
Pipeline	179	59	59
Real Image	296	2,321	1,910
Stat./Analysis	313	8,756	7,945
Total	**4,844**	**1,56,184**	

Table 12: AI-FIGURES Dataset Across Categories and Cleaning Stages

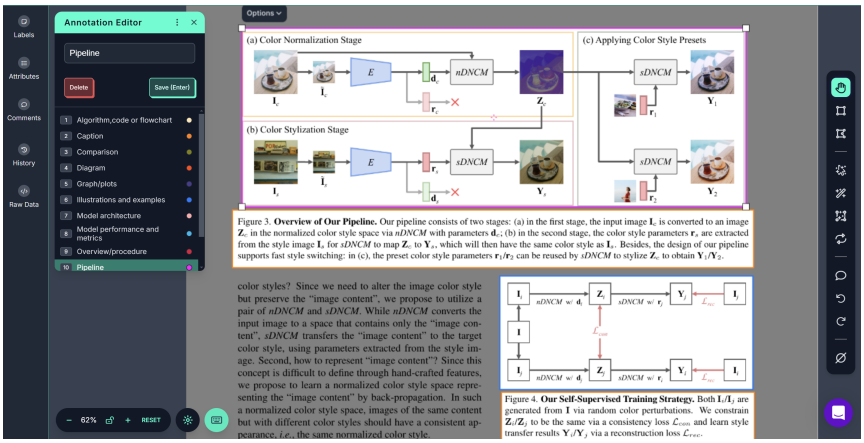


Figure 6: Roboflow Annotate Platform

- Do not include surrounding text or unrelated parts of the page in the bounding box.
- Do not annotate tables or equations; this task is only for figures and captions.
- There should be no overlapping or duplicate annotations.
- Class Definitions: Each figure must be assigned exactly one of the following classes:
  - **Caption**: Text that describes a figure.
    Example: "Figure 3: Architecture of the proposed model."
  - **Diagrams**: Schematic representations, flowcharts, or conceptual illustrations.
    Examples: System design diagrams, logic flow representations.
  - **Graphs/Plots**: Graphs, charts and mathematical plots.
    Examples: Line graphs, bar charts, scatter plots, histograms.
  - **Illustrations and Examples**: Figures providing explanatory visual aids for a concept or process.
    Examples: Illustrative sketches, educational examples, artistic depictions.
  - **Model Architecture**: Figures depicting the structural design of machine learning or deep learning models.
    Examples: Transformer model, LSTM or YOLO architectural diagram
  - **Statistics and Analysis**: Figures containing statistical results, experimental comparisons, or analytical visualizations.
    Examples: Performance comparison graphs, confusion matrices, regression analysis plots.
  - **Overview/Procedure**: Figures illustrating a high level overview of multi-step processes, workflows, or methodologies without detailed representations of each component.
    Examples: Overview of the object detection process using YOLO
  - **Pipeline**: Figures representing entire processing workflows, often spanning multiple steps and modules.
    Examples: End-to-end ML pipeline diagrams

- **Model Performance and Metrics**: Figures showing model evaluation results, benchmarking, and performance graphs.
  Examples: Precision-recall curves, accuracy vs. epochs graphs, performance tables.
- **Real Images**: Photographic images or realistic visual content extracted from real-world sources.
  Examples: Images from datasets, captured photographs, images of people, animals, places or objects.
- **Algorithms/Code/Flowchart**: Figures containing algorithmic representations, such as code snippets, pseudo code, or flowcharts.
  Examples: Code blocks (e.g., Python, C++, pseudo code), Flowcharts detailing algorithmic steps, Structured representations of an algorithm's execution flow.

C Dataset Construction - Manual Evaluation

C.1 Annotation Guidelines

For each figure, its corresponding caption, class category and a link to the original pdf from which the figure was extracted is provided. The evaluator must :

- Choose "Yes" under "Acceptable" if the figure is segmented and classified correct and paired with the correct caption as per the parent paper pdf supplied.
- If "No" is selected, then the issue must be narrowed down to one of the following cases:
  - Choose "Figure Segmentation Error" if the figure is cropped in a wrong fashion.
  - Choose "Figure Classification Error" if the figure is classified into the wrong category.
  - Choose "Figure-Caption Pairing Error" if the figure is paired with the wrong caption.

D Captioning - Manual Evaluation

D.1 Annotation Guidelines

For each model, evaluate whether the caption generated provides a comprehensive description of its figure. An exact match is not expected with the ground truth caption, but there must be some degree of alignment in the content.

- If the caption generated by a particular model is acceptable, select "Yes".
- If you have selected "No" then narrow down the issue to one of the following:
  - Oversimplification: The oversimplified caption is too short compared with the original ground truth caption.
  - Contextual misunderstanding : Contextual Misinterpretation refers to captions with unmentioned content in the figure.
  - Recognition Error : Recognition Error denotes the model wrongly identified the number or text in the figure.

E Model Description

E.1 Models for Figure Captioning

Molmo-7B: Molmo-7B-D-0924 (Deitke et al., 2024) is a multimodal AI model developed by the Allen Institute for AI, designed to integrate vision and language understanding. Built upon the Qwen2-7B architecture and utilizing OpenAI's CLIP as its vision backbone, this model has been trained on PixMo, a curated dataset of 1 million image-text pairs. Molmo-7B-D-0924 achieves an average score of 77.3% across 11 academic benchmarks and holds a human preference Elo rating of 1056, positioning its performance between GPT-4V and GPT-4o. The model is fully open-source, with all associated artifacts, including the PixMo dataset, training code, evaluations, and intermediate checkpoints, available to the public.

InternVL2_5-8B: This is a multimodal large language model developed as part of their InternVL 2.5 series. This model integrates a vision component, InternViT-300M-448px-V2_5, with a language component, InternLM2_5-7B-Chat, connected through a randomly initialized MLP projector. The architecture follows the "ViT-MLP-LLM" paradigm, employing a pixel unshuffle operation to reduce the number of visual tokens and a dynamic high-resolution strategy to handle various data types, including single images, multiple images, and videos. The training process is structured across three stages: MLP warmup, contrastive learning, and generative learning, aiming to enhance the model's visual perception and multimodal capabilities. InternVL2_5-8B (Chen et al.,

2024) has demonstrated proficiency in tasks such as multimodal reasoning, OCR, chart and document understanding, and video comprehension.

Qwen2-VL-7B-Instruct: This is an advanced vision-language model developed by Qwen, designed to handle a variety of visual and textual tasks. This model supports arbitrary image resolutions, dynamically converting them into visual tokens for more human-like visual processing. Qwen2-VL (Wang et al., 2024) achieves state-of-the-art performance on visual understanding benchmarks, including MathVista, DocVQA, RealWorldQA, MTVQA, etc. Additionally, it offers multilingual support, understanding texts in languages such as English, Chinese, most European languages, Japanese, Korean, Arabic, and Vietnamese.

MiniCPM-V: MiniCPM-V (Yao et al., 2024) is a multimodal large language model designed for deployment on devices ranging from GPU cards to mobile phones. By compressing image representations into 64 tokens via a perceiver resampler, it achieves high efficiency with reduced memory usage and faster inference speeds. Despite its compact size of 3 billion parameters, MiniCPM-V demonstrates state-of-the-art performance on multiple benchmarks, surpassing existing models of comparable size and even rivaling larger models like Qwen-VL-Chat. Notably, it supports bilingual multimodal interactions in English and Chinese, making it versatile for diverse applications.

Janus-Pro-7B: Janus-Pro-7B (Chen et al., 2025) is an advanced multimodal AI model developed by DeepSeek, designed to unify text and image processing capabilities within a single framework. In text-to-image tasks, Janus-Pro-7B excels in generating high-quality images from textual descriptions, outperforming models like OpenAI's DALL-E 3 and Stability AI's Stable Diffusion in various benchmarks. For image-to-text tasks, Janus-Pro-7B employs a decoupled visual encoding approach, utilizing the SigLIP-L vision encoder to process images at resolutions up to 384x384 pixels. This design allows the model to effectively understand and generate textual descriptions of visual content, making it versatile for applications requiring both image generation and comprehension.

GIT (Base and Large): GIT (Generative Image-to-Text) (Wang et al., 2022) is a Transformer-based model developed by Microsoft for vision-language tasks such as image and video captioning, visual question answering (VQA), and image classification. The model is conditioned on both CLIP image tokens and text tokens, enabling it to generate textual descriptions based on visual inputs. GIT is available in two primary configurations:

- GIT-Base: This version comprises approximately 177 million parameters and is trained on 10 million image-text pairs.
- GIT-Large: This larger variant contains around 395 million parameters and is trained on 20 million image-text pairs. The expanded parameter count enhances its capacity to generate more detailed and accurate textual descriptions from images, making it well-suited for complex vision-language tasks.

Both versions utilize a Transformer decoder architecture, where the model has full bidirectional attention over image patch tokens and causal attention over text tokens. This design enables the models to predict the next text token by considering both the visual input and the preceding text, facilitating coherent and contextually relevant text generation based on images.

E.2 Tag Classification

Llama 3.2-1B Instruct: The Llama 3.2 collection of multilingual large language models (LLMs) (Grattafiori et al., 2024) is a collection of pretrained and instruction-tuned generative models in 1B and 3B sizes (text in/text out). The Llama 3.2 instruction-tuned text only models are optimized for multilingual dialogue use cases, including agentic retrieval and summarization tasks. They outperform many of the available open source and closed chat models on common industry benchmarks.

Llama-3.1-8B-Instruct: Llama-3.1-8B-Instruct (Grattafiori et al., 2024) is an 8-billion-parameter language model developed by Meta as part of the Llama 3.1 series, released in July 2024. This model is fine-tuned for instruction-based tasks, enhancing its performance in understanding and generating human-like text responses. It supports eight languages: English, German, French, Italian, Portuguese, Hindi, Spanish, and Thai. Notably, Llama-3.1-8B-Instruct features an expanded context window of up to 128,000 tokens, allowing it to process and generate longer sequences of text effectively.

Mistral 7B Instruct v0.2: Mistral-7B-Instruct-v0.2 (Jiang et al., 2023) is an instruction fine-tuned version of the Mistral-7B-v0.2 language model, developed by Mistral AI. This iteration introduces

key improvements over its predecessor, including an expanded context window of 32,000 tokens (up from 8,000), a RoPE-theta value of 1e6, and the removal of Sliding-Window Attention. These enhancements enable the model to generate coherent and contextually rich responses, making it suitable for a wide range of natural language processing tasks.

Qwen2.5-0.5B-Instruct and Qwen2.5-7B-Instruct: Qwen2.5-0.5B-Instruct (Team, 2024) is a 0.5 billion parameter instruction-tuned language model developed by the Qwen team at Alibaba Cloud. As part of the Qwen2.5 series, this model offers significant improvements in instruction following, coding, mathematics, and multilingual support across over 29 languages, including Chinese, English, French, and Spanish. It features a context length of up to 32,768 tokens and can generate sequences up to 8,192 tokens.

F Task Prompts

In this section we provide the prompts that we have used for the various tasks in our study. Each prompt is designed to guide the model in performing specific operations, ensuring clarity, coherence, and consistency in the generated outputs. In a task, the same prompt is used for all models under comparative evaluation. Below, we list the prompts used under each task.

F.1 Figure Captioning

Zeroshot Captioning:

"Generate a concise and articulate caption for a diagram retrieved from a research paper. Focus on explaining the key idea or concept represented by the diagram in no more than 100 words. Avoid describing the structural elements or layout of the diagram, and ensure the caption is self-contained and conceptually meaningful without external references."

Captioning using paper title as context:

"Using the context provided in the following title: <title>, generate a concise and meaningful caption for the image that explains the key concept or core idea represented by the figure in no more than 100 words."

Captioning using paper title and abstract as context:

context = " Title: <title>
Abstract: <abstract>"

"Using the context provided below, generate a concise and meaningful caption for the image that explains the key concept or core idea represented by the figure in no more than 100 words: <context>"

F.2 Tag Classification

"You will receive a figure caption and a list of predefined figure categories. Your task is to classify the caption into exactly one category based on the concept it represents. If the caption aligns with multiple categories, choose the most appropriate one that best describes the figure type.
***Categories**:*
- Diagram: schematic figures or sketches.
- Graphs-plots: charts and plots.
- Illustrations and examples: figures providing examples or visual aids.
- Model architecture: figures depicting the architecture of models.
- Statistics and Analysis: figures or graphs involving statistical results and analysis.
- Overview-procedure: figures that illustrate a high level overview of methods or procedures.
- Pipeline: figures showing complete workflows.
- Model performance and metrics: figures or graphs showing performance evaluation of models.
- Real image: photographs or realistic images.
***Examples**:*
Example 1:
Caption: A scatter plot showing the relationship between training time and model accuracy, with a trend line fitted to the data.
*Your response: **Graph-plots***

Example 2:
Caption: A step-by-step workflow illustrating the data preprocessing, model training, and evaluation stages in a deep learning pipeline.
*Your response: **Pipeline***
***Instructions**:*
- Identify the most relevant category for the caption.
*- The classification must reflect **only one category**, avoiding overlaps. If multiple categories seem relevant, choose the broadest and most appropriate one*
- Return only the category name. Do not add extra explanation, reasoning, or special characters to your response.
- Return the exact category name as it appears in the list without any variations
***Figure Caption** :*
<caption>

Your Response:

F.3 Generating QA pairs using InternVL

<image>
<caption>
Using the visual content of this image and the context provided by the caption, generate 3 simple and self-contained question-answer pairs.
Ensure that:
1. The questions are directly answerable using the content of the image and/or the caption.
2. The questions are straightforward and do not require multi-step reasoning.
3. The answers are contained entirely within the image and caption.
4. The questions do not point to any external references.
Provide the output in the format:
Q1: ...
A1: ...
Q2: ...
A2: ...

F.4 Finetuning Qwen for Question-Answering

You are a Vision Language Model specialized in interpreting visual data from graphs, charts and figures depicting statistical analysis. Your task is to analyze the provided figure and respond to queries with concise and informative answers, usually in one or two sentences. Focus on delivering accurate, succinct answers based on the visual information. Avoid additional explanation unless absolutely necessary.

F.5 Using finetuned Qwen for question-answering on MathVista

Provide a clear, short, and succinct numerical answer to the question based entirely on the given figure, without any external references or extra words.Follow the hint given below closely:
<hint>
***Question**:*
<question>
***Answer**:*

F.6 Locality-Specific Question Response Generation on M3SciQA

You are given a figure, a question, and a list of paper candidates of titles and abstracts. Your task is to answer the question based on the figure information, then order the paper candidates that I provide to you so that the paper that is more relevant to the question comes first in the list. Return a minimum of 1 and a maximum of 5 paper candidates in the rank list. Ideally there should be 3 paper candidates.
Provide your answer at the end in a json file of this format using S2_id only:{{"ranking":[rank_1_s2_id, rank_2_s2_id] }}.
Make sure the responded list is in a valid format and that it only contains the S2_id. Do not include the title or abstract in the answer list. Also report the s2 ids in a comma separated manner.
<question>
{question}
</question>
<paper candidates>
{reference_title_abstract_list}
</paper candidates>

G Tag Classification: Confusion Matrix Evaluation

We present confusion matrices for each model used in Tag classification task. These confusion matrices illustrate the distribution of predicted tags against the actual tags, highlighting patterns of correct and incorrect classifications.
By analyzing these matrices, we can identify common misclassifications and assess how well each model distinguishes between different tag categories.

H Human Evaluation Platforms

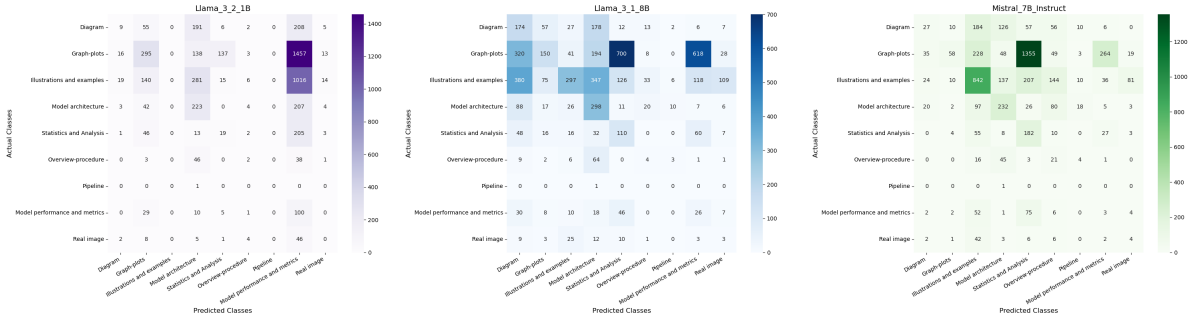
H.1 Evaluation of Figure Extraction and Classifcation in AI Figures

We present a view of the interface that was used by our evaluators for assessing the quality of figures and captions present in our dataset.

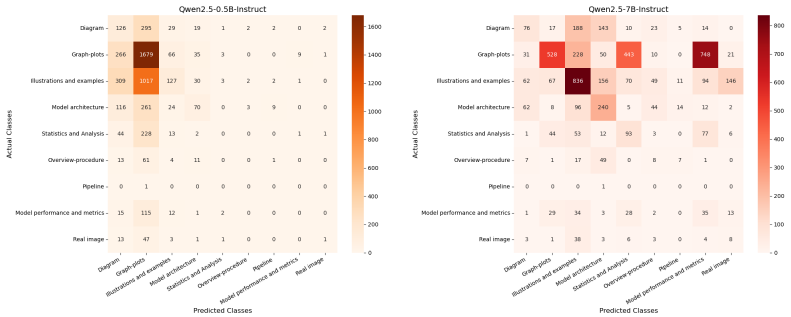
H.2 Evaluation of Figure Captioning

We present a view of the interface that was used by our evaluators for assessing the quality of captions generated by each model under evaluation.

I Finetuned SDXL Figure Generation results



(a) Llama-3.2-1B (b) Llama-3.1-8B (c) Mistral-7B-v0.2



(d) Qwen2.5-0.5B (e) Qwen2.5-7B

Figure 7: Confusion Matrix for Tag Classification

Figure 8: Human Evaluation of AI Figures

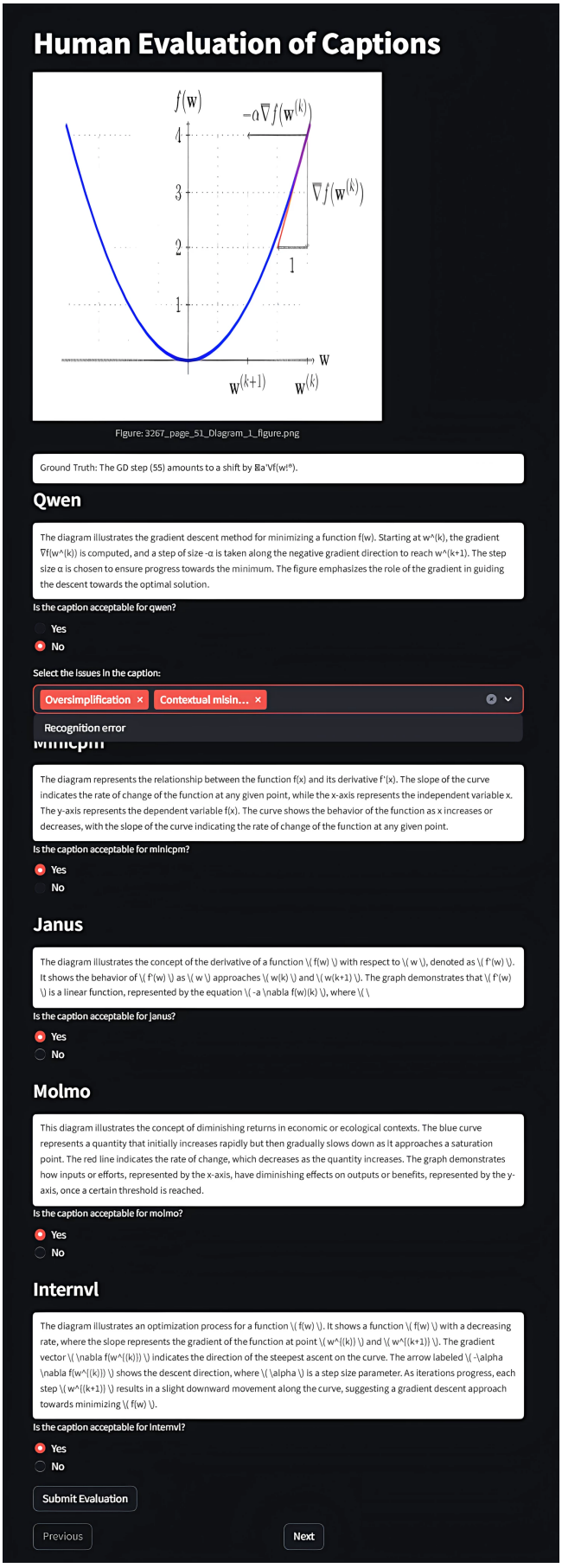
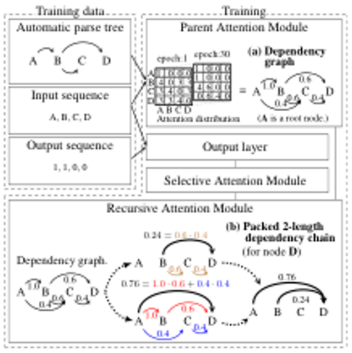
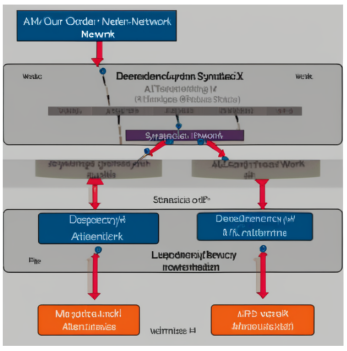


Figure 9: Human Evaluation of Captioning Results

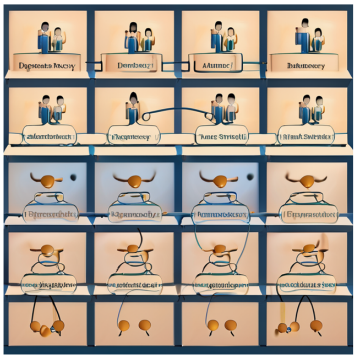
Original figure



Finetuned SDXL



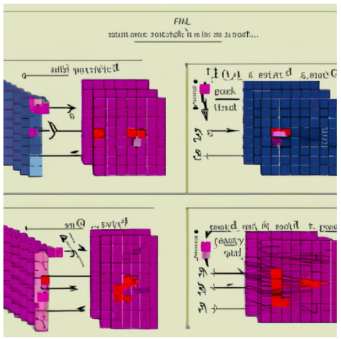
Zero-shot SDXL



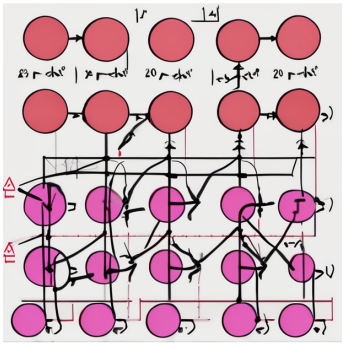
Original Caption:
Dependency structures used in our higher-order syntactic attention network.

Original figure

Finetuned SDXL

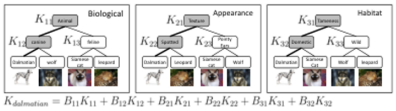


Zero-shot SDXL

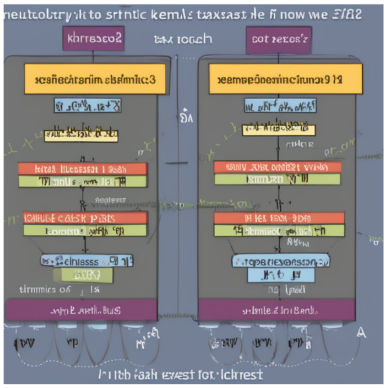


Original Caption:
A proof script in Cog (left) and the resulting proof states, proof steps, and the complete proof tree (right).

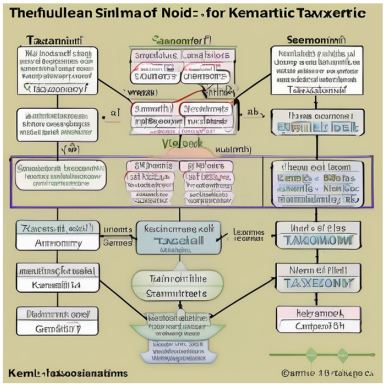
Original figure



Finetuned SDXL



Zero-shot SDXL



Original Caption:
Main idea: For a given set of classes, we assume multiple semantic taxonomies exist, each one representing a different view of the inter-class semantic relationships. Rather than commit to a single taxonomy which may or may not align well with discriminative visual features we learn a tree of kernels for each taxonomy that captures the granularity-specific similarity at each node. Then we show how to exploit the inter-taxonomic structure when learning a combination of these kernels from multiple taxonomies (i.e., a kemel forest) to best serve the object recognition tasks.

Figure 10: A comparative analysis of figure generations by the fine-tuned SDXL model, with the original figure and zero-shot generations from the base SDXL model.