Diagnostics-LLaVA: A Visual Language Model for Domain-Specific Diagnostics of Equipment

Aman Kumar, Mahbubul Alam, Ahmed Farahat, Maheshjabu Somineni and Chetan Gupta

Industrial AI Lab, Research & Development, Hitachi America Ltd., Santa Clara, CA, 95054, USA
aman.kumar@hal.hitachi.com
mahbubul.alam@hal.hitachi.com
ahmed.farahat@hal.hitachi.com
maheshjabu.somineni@hal.hitachi.com
chetan.gupta@hal.hitachi.com

Abstract

The recent advancements in the area of Large language models (LLMs) have opened horizons for conversational assistant-based intelligent models capable of interpreting images, and providing textual response, also known as Visual language models (VLMs). These models can assist equipment operators and maintenance technicians in complex Prognostics and Health Management (PHM) tasks such as diagnostics of faults, root cause analysis, and repair recommendations. Significant open-source contributions in the area of VLMs have been made. However, models trained in general data fail to perform well in complex tasks in specialized domains such as diagnostics and the repair of industrial equipment. Therefore, in this paper, we discuss our work on the development of Diagnostics-LLaVA, a VLM suitable for interpreting images of specific industrial equipment, and provide better response than existing open source models in PHM tasks such as fault diagnostics and repair recommendation. We introduce Diagnostics-LLaVA based on the architecture of LLaVA and created one instance of Diagnostics-LLaVA for the automotive repair domain, referred to as Automotive-LLaVA. We demonstrate that our proposed Automotive-LLaVA model performs better than the state-of-the-art open-source visual language models such as mPlugOWL and LLaVA in both qualitative and quantitative experiments.

1. Introduction

The development of domain-specific visual language models has emerged as an important area of research due to the increasing demand for advanced artificial intelligence systems that can communicate, reason, and understand the visual world effectively (Park & Kim, 2023). A Visual Language Model (VLM) combines the capabilities of Computer Vision (CV) and Natural Language Processing (NLP) to create a system that comprehends and generates descriptions based on visual content with the help of large language models (LLMs) (Wang et al., 2023). Within the field of prognostics and health management (PHM), a domain-specific VLM tailored to the needs of equipment operators and maintenance technicians has the potential to revolutionize the maintenance and repair of equipment in various industries (Lai et al., 2024). By leveraging a domain-specific VLM, operators and technicians can seamlessly interact with such intelligent systems, which can automatically analyze equipment components, identify issues, and communicate relevant information in an efficient and intuitive manner. As technology continues to advance, such a specialized VLM will enable technicians to streamline diagnosis and repair processes, increase operations and maintenance efficiency, and ultimately enhance overall user satisfaction and safety.

Recent advancements in Visual Language Models (VLMs) have significantly improved the integration of computer vision and natural language processing (He et al., 2024). Notable developments include the Multi-modal Instruction Tuned LLMs with Fine-Grained Visual Perception (AnyRef) model which generates pixel-wise object perceptions and natural language descriptions from multi-modality references (X. Zhao et al., 2024). Additionally, the LLaVA model (Liu, Li, Wu, & Lee, 2024) enhances visual processing by integrating multi-granularity images and introducing a novel visual instruction tuning method for extending MLLMs to perform various multi-modal tasks, surpassing previous state-of-the-art performance on multiple visual instruction tuning benchmarks. mPLUG-Owl (Ye et al., 2023) is another popular open-source VLM. mPLUG-Owl2 (Ye et al., 2024), an extension of the mPLUG-Owl model, revolutionizes multi-modal large language models by effectively leveraging modality collaboration to improve performance in both text and multi-modal tasks. Despite these advancements, some VLMs do not align with human vision illusions, particularly for question-answering tasks (Y. Zhang, Pan, Zhou, Pan, & Chai, 2023). Interestingly, research from MIT's CSAIL demonstrates that language models trained on text possess a strong understanding of the visual world, opening up possibilities for improved multi-modal understanding and interaction between humans and AI systems (MIT, 2024).

The capabilities of large language models extend beyond conversation; they can also be used to query information from documents through Retrieval Augmented Generation (RAG) (Lewis et al., 2020). A recent study utilized automotive manuals for information retrieval, highlighting the limitation of interpreting visual elements (Medeiros et al., 2023). This underscores the need for a VLM that can enhance knowledge retrieval by effectively processing both textual and visual information in specialized domains (Kumar & Starly, 2022). RAG can be highly suitable for cases where enterprise information needs to be retrieved from highly classified documents that should remain on-premises (R. Zhao et al., 2023). Domain-specific VLMs can help achieve better results due to their advanced image processing capabilities.

Domain-specific VLMs have led to the development of innovative techniques for transforming and extending these models to cater to unique industry requirements (J. Zhang, Huang, Jin, & Lu, 2024). By applying methods that transform domain-specific visual and vision-language datasets into a unified question-answering format, such as Visual Question Answering Instruction (VQA-IN), researchers have successfully extended Multimodal Large Language Models (MLLMs) to domain-specific tasks (Lee, Cha, Lee, & Yang, 2024; Vidyaratne et al., 2024). These advancements enable the models to achieve high performance on domain-specific visual tasks while maintaining their performance on vision-language tasks in a multi-task manner, demonstrating the potential for domain-specific VLMs to revolutionize various industries and facilitate seamless multi-modal interactions between humans and AI systems. Therefore, a domain-specific VLM designed for a specific industry could help operators and technicians in that industry analyze equipment components and identify issues efficiently, ultimately enhancing diagnosis and repair processes.

In this paper, we propose a method for building Diagnostics-LLaVA, a VLM that is suitable for assisting in maintenance and repair tasks in the specialized industrial domain and is based on the architecture of the LLaVA model. We developed Automotive-LLaVA, an instance of Diagnostics-LLaVA for the automotive domain. The choice of the automotive domain is motivated by the availability of large enough datasets of public documents in the automotive domain that can be used for development and validation. However, the proposed method can be applied to other industrial domains with specialized proprietary documents.

The development of Automotive-LLaVA involves the following steps.

1. We construct the Automotive image-text pairs dataset for training the Contrastive Language-Image Pretraining (CLIP) based Vision Transformer (ViT) model for joint text and image representation through adapters (Radford et al., 2021; Dosovitskiy, 2020).

2. We create an Automotive Visual Question Answering (AVQA) dataset with images, questions and responses for supervised instruction tuning of the LLaVA model.

3. We propose a benchmark for the evaluation of the Automotive-LLaVA model.

The remainder of the paper is organized as follows. Section 2 describes the process of data preparation for the VLM training. Section 3 explains the architecture and the model training process. Section 4 presents the quantitative and qualitative experiments to evaluate the models. Sections 5 and 6 talk about the conclusion, limitations, and future work.

2. DATA PREPARATION

2.1. Domain data collection

Acquiring domain-specific data poses significant challenges but is crucial for effective training in large language models. While text data acquisition is relatively straightforward, collecting multi-modal image-text pair data is more complex due to the essential relationship between image and text pairs required for multi-modal model training. To facilitate this, we collect data from trusted sources of information in the domain such as operation and maintenance manuals, where corresponding figures and text are readily available, enabling the creation of robust image-text pairs.

The data preparation was done in four stages: (1) Extraction of images from the domain-specific documents (mostly in PDF format) (2) Extraction of image-text pairs dataset for pretraining (3) Building domain-specific instructions dataset for visual question answering (4) Data Augmentation

2.1.1. Extraction of images from domain-specific PDF documents

We use PyMuPDF python library (Guedes & da Silva, 2021) to parse the PDF document. We created a function to locate the area occupied by the image caption and the corresponding image. Next, we retrieve all blocks positioned above and to the right of the image caption area. We group these elements using a threshold method to distinguish between distinct elements and exclude any groups identified as text bodies, focusing solely on those likely to contain images.

2.1.2. Building image-text pairs dataset for pretraining

We focus on extracting all images from different parts of the PDF. Each image in the PDF is accompanied by a caption that provides brief information about the image. We collect these captions, whether they are above or below the image, referring to them as short descriptions. Additionally, we extract all text from paragraphs where the figure number is mentioned, which we call long descriptions. This ensures that we not only have brief captions but also more detailed descriptions, which are very useful during the later stages of the work.

For the text part, we leverage the manual's textual content and GPT-4's (Achiam et al., 2023) capabilities to create the final data. We use the GPT-4 to rephrase the short descriptions so that they align well with our format and remove unnecessary phrases such as ”Figure XXX”. For the long descriptions, we observe that not every line mentioning the figure number is useful, and sometimes the paragraphs are quite lengthy. Therefore, we perform similar operations as with the short descriptions, including removing irrelevant information by appropriately prompting GPT-4. In total, for the automotive domain, we collected 3,287 unique image-text pairs.

2.1.3. Building Instructions dataset for visual question answering

To better understand the input and generate relevant responses following instructions, Vision-Language Models (VLMs) are typically trained using a technique called instruction tuning. This involves training the model to align with desired behaviors or tasks provided through explicit instructions. The model is given examples containing both the input and the desired output. To adapt VLMs for domain-specific tasks such as automotive, an instruction dataset specialized in the automotive domain is required. As of the time of writing this paper, there is no existing instruction dataset available for the automotive domain. To address this gap, we created the Automotive Visual Question Answering (AVQA) dataset.

For each image, we created pairs of questions and answers to train a visual QA model. Using the image and text dataset developed in the previous step, we employed the advanced GPT-4V model to read both the image and text and generate five questions per image following specific instructions provided as prompts. Before generating the AVQA data, we applied simple rules to filter images based on their descriptions to ensure they were suitable for question generation. Ultimately, 2,477 images were selected for the AVQA dataset. This was followed by data augmentation, which is explained in the next section.

2.1.4. Data Augmentation

All the automotive images used for training the image-text representation model and the instruction-tuned model were augmented using the following techniques: rotation, blur, grayscale, edge detection, and sharpening (Shorten & Khoshgoftaar, 2019). This augmentation increased the dataset by a factor of six. As a result, we obtained 19,722 images for image-text pairs from the initial 3,287 unique images (A) and 14,682 images for visual QA pairs from the initial 2,477 unique images (B), where B is a subset of A.

The authors of LLaVA have curated the dataset in such a way that each image is accompanied by one description per prompt. For our dataset curation, we decided to augment this by providing four additional types of descriptions with different responses based on our acquisition of both short and long descriptions of each image, as mentioned in section 2.1.2. We selected a variety of prompts already present in the original LLaVA dataset. An example of pretraining and finetuning instructions for a given image has been provided in the Appendix. The instruction prompts used for automotive image-text pairs are as follows:

1. Describe the image concisely: For this prompt, we use the image caption as provided in the document.

2. Provide a detailed description of the image: We compile all occurrences of the figure number in the document to create a detailed description of the image with the help of GPT-4.

3. Share an informative description of the image: We use a 2-3 line summary of the detailed description.

4. Give a brief summary of the image: We provide a one-line summary of the detailed description.

To sum, 19,722 images when augmented with four descriptions, resulted in a total of 78,888 image-text pairs from 3,287 unique images.

2.1.5. Mixing of general and domain-specific data

We mix the general dataset curated by the authors of LLaVA-1.5 with the domain-specific data to perform pretraining and instruction tuning of the model from scratch. For pretraining, we used the LLaVA Visual Instruct Pretrain LCS-558K dataset, which contains 558K general image-text pair examples and utilized LLAVA 665K instruction-following examples (we call this dataset D1) (Y. Zhang, Zhang, et al., 2023). For the automotive domain, we added 79K automotive image-text pair examples, resulting in a total of 637K examples for pretraining. For instruction tuning, we added 15K automotive instruction examples, bringing the total to 680K examples (we call this dataset D2). Table 1 shows the D1 and D2 datasets used in the VLM training experiments discussed in the subsequent sections of the paper.

2.2. Evaluation dataset

For the automotive domain, the Automotive Service Excellence (ASE) certifications (Kolo, 2006) are designed to gauge

Table 1. Dataset

Dataset name	pretraining data size	finetuning data size
D1	558K	665K
D2	637K	680K

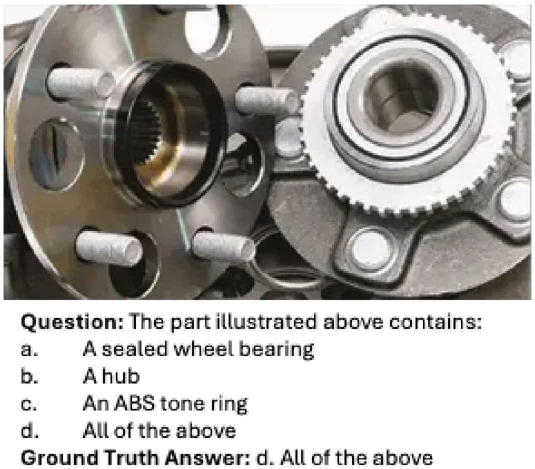


Figure 1. Example of Automotive Service Excellence exam question. (freeasestudyguides, 2024)

automotive service professionals' expertise in vehicle repair, service, and parts distribution. ASE exams cover various sections that target different specialties within the automotive sector. We selected this exam since it serves as an excellent benchmark for evaluating knowledge in the automotive domain (Yemaneab, 1997).

We acquired sample exams and curated an evaluation dataset by parsing the PDFs. The sample exam contains 1,090 questions from 10 categories, of which 876 are text-based and 214 are image+text-based. In this work, we focus on all 214 image+text-based questions. Each question is a multiple-choice question (MCQ) with four options and an associated image. The models are prompted to answer the question with one correct option, and the results are compared against the ground truth labels. An example of an ASE exam question is given in Fig. 1.

3. VLM MODEL TRAINING

3.1. Model training

The Automotive-LLaVA architecture is designed to enhance visual and language processing capabilities in the automotive domain. This architecture, illustrated in Fig. 2, includes a pretrained visual backbone for image encoding, a chat-based large language model for generating responses, and a projection network that links the visual backbone to the language model (Liu et al., 2024).

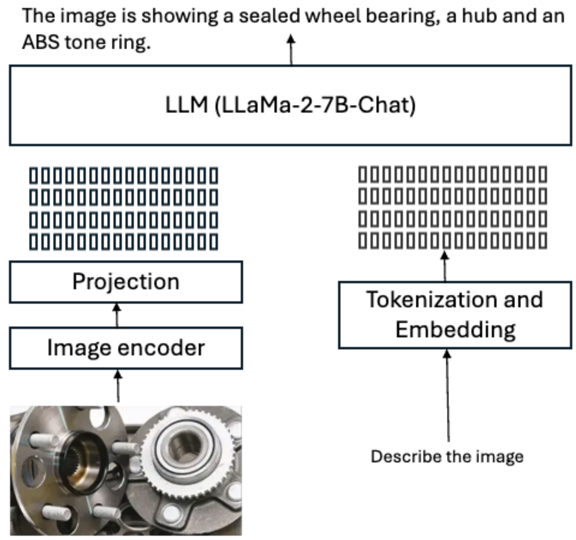


Figure 2. Architecture of Automotive-LLaVa.

Consider a sample $\{X, I, T\}$ from the Automotive-instructions dataset, with X representing the image, I as the instruction, and T as the instruction's response. Initially, an image encoder processes the input image $X \in R^{H \times W}$, where H, and W, denote the image's height, and width, respectively. This encoder transforms the image into a series of image tokens $Z_X \in R^{N \times D_{img}}$, with N indicating the token sequence length and D_{img} the image encoder's dimensional capacity.

Following this, the tokens pass through a projection network. This network projects the visual tokens into an embedding space of dimension D_{emb}, producing the mapped sequence $F_{img} \in R^{N \times D_{emb}}$. These image features are subsequently combined with the instruction tokens $F_{inst} \in R^{M \times D_{emb}}$. The concatenated sequence forms the input $F \in R^{K \times D_{emb}}$ for the LLM, where $K = M + N$.

The LLM is a chat-based language model employing the transformer architecture. It receives the sequence F consisting of both visual and linguistic tokens as input and generates the response in an auto-regressive manner. This process involves optimizing the probability distribution for generating an accurate response given the combined image-instruction tokens. The probability distribution can be expressed as follows:

$$P(T|X, I) = \prod_{k=1}^{K} P(T_k | T_1, \ldots, T_{k-1}, I, X)$$

where K denotes the length of the response sequence, and $P(T_k | T_1, \ldots, T_{k-1}, I, X)$ represents the probability of the k-th token given the preceding tokens, the instruction, and the image.

For our experiments we have undergone two stages of train-

Table 2. Pretraining experiments showing different LLM and the dataset used with a ViT-L/14 Visual encoder

Experiment	LLM	Data
LLaVa-1.5	Vicuna-1.5-7B	D1
LLaVa-LLaMa (ours)	LlaMa-2-7B-chat	D1
LLaVa-Auto (ours)	LlaMa-2-7B-chat	D2

ing: (1) pretraining for feature alignment and (2) Visual Instruction tuning.

3.2. Feature Alignment and Visual Instruction tuning

As discussed in the data preparation section, the pretraining data consists of a mix of general and automotive data. Each sample in the dataset can be treated as a single-turn conversation between input and assistant where instruction is provided to the assistant to explain the image. The ground truth prediction answer is provided which is used during the training process. Keeping the visual encoder and LLM weights frozen, we train the model to maximize the likelihood of trainable parameters of the projection matrix. Through this, the image features F_{img} can be aligned with the LLM embedding using image-text pairs.

In the Visual Instruction tuning step, we keep the visual encoder weights, and the pretrained weights of the projection layer frozen and finetune the LLM weights using a visual question-answering dataset as discussed earlier.

4. Experimental Results

4.1. Experimental Settings

Automotive-LLaVa model uses a mix of automotive + general data for pretraining and finetuning, and LLaMa-2-7b-chat model (Touvron et al., 2023) as the LLM while basing its architecture on LLaVa.

As shown in Table 2, Experiment E1 is the original LLaVa model which used Vicuna-1.5-7B chat LLM and D1 dataset (refer to section 2.1.5), while E2 and E3 were created by us where we experimented by replacing Vicuna LLM with LLama-2-7B-chat model in E2 and then replacing the original dataset (D1) with newly created dataset (D2) in E3.

We train all our models with 8 $\times$ H100s GPUs using the parameters provided in LLaVA. We pretrain our models for 1 epoch with a learning rate of 2e-3 and a batch size of 128 and finetune for 3 epochs with a learning rate of 2e-5 and a batch size of 32, which took 2.5 hours for pretraining and 4.75 hours for finetuning on 8 $\times$ H100s.

4.2. Quantitative Evaluation

The models created upon training were used to perform evaluation inference on the ASE benchmark dataset. VLMs in general are good for free-form responses to explain the image and ask questions about the same. However, due to the limited capabilities of a small-sized model, it is difficult to expect consistent output in the desired format. We performed some iterations of prompting the VLM to produce the output as one of the options provided in zero-shot and few-shots (5-shot) (Brown et al., 2020). Providing few-shot image and text pairs to VLM as input and expecting the final question response based on another image would require additional steps which were beyond the scope of this work, therefore we focussed on experimenting with zero-shot setting. We experimented with 5-shot text examples, but it did not change the way the model provided output so we stuck to zero-shot for the final results. We prompted the model by providing the same question and image three times to note the variability in the output. Due to inconsistent output by VLM, we employed GPT-4 to check for the inconsistency and convert the response into one of the 4 options so that it can be validated directly against the ground truth.

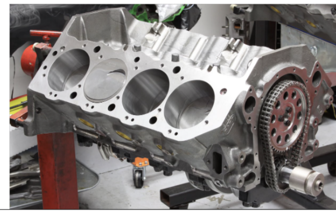
Table 3. Results on a total of 214 ASE Questions

VLM Model	Correct	Accuracy
mPlugOWL	44	20.56
LLaVa-1.5	62	28.97
LLaVa-v1.6-mistral-7b	60	28.04
LLaVa-v1.6-vicuna-7b	52	24.29
LLaVa-LLaMa (ours)	71	33.18
LLaVa-Automotive (ours)	67	31.31

The results of the evaluation of all the models are presented in the table 3. We show that the LLaVA-LLaMA model has outperformed all open source models in comparison by more than 4%. This depicts that the base LLM Vicuna used in LLaVA is not very suitable for domains such as automotive. Replacing the LLM with LLaMA-2-7B-chat resulted in significant improvement in the benchmark. We also show that the dataset created for training a LLaVA-automotive model when used with LLaMA-2-7B-chat does not yield better results, on the contrary, it reduces the performance. However, the result achieved through LLaVA-automotive is still more than 2% better than the original LLaVA. More investigation is required on the incorrect answers to gauge what is hurting the performance of LLaVA-Automotive. Moreover, from this observation, we decided to do an investigation on the qualitative aspect to evaluate the response provided by the LLaVA-Automotive model. The results are shown in the next section.

4.3. Qualitative Evaluation

Towards the qualitative evaluation, we selected random images related to automotive from the web and prompted the three models i.e. LLaVA-1.5, LLaVA-LLaMA and LLaVA-Automotive to analyze the image and answer the question. The same images and the questions were fed to GPT-4 and prompted to rate the three model responses on a scale of 10. GPT-4 provided both the pros and the cons of the model re-



User	Analyze the image carefully, think step by step and provide all possible components that you can recognize from the image.	GPT4 rating
LLaVA-1.5	The image features a close-up view of a large engine block, likely a V8 engine. The engine block is made of metal and has a chrome finish. The engine has a total of six cylinders, with each cylinder having a distinct shape and size. The cylinders are arranged in a row, with one cylinder on the left side, two cylinders in the middle, and three cylinders on the right side of the engine. The engine block is the central focus of the image, showcasing its intricate design and craftsmanship.	5
LLaVA-LLaMA	When analyzing the image, I can recognize the following components of the engine: 1. Cylinder: The engine has multiple cylinders, which are the chambers where the combustion process takes place. 2. Piston: The piston is a crucial component of the engine, responsible for moving back and forth inside the cylinder to convert the linear motion of the combustion process into rotational motion. 3. Cylinder head: The cylinder head is the top part of the cylinder, which houses the valves and other components that control the flow of fuel and exhaust gases. 4. Valves: The valves are the components that control the flow of fuel and exhaust gases into and out of the cylinder. They are typically located on the cylinder head and are opened and closed by the piston's movement. 5. Crankshaft: The crankshaft is the component that converts the linear motion of the piston into rotational motion, which is then transmitted to the wheels of the vehicle. These components work together to create the engine's power-generating mechanism, allowing the vehicle to move forward.	4
LLaVA-Automotive	The image features a close-up view of a cylinder block, which is a part of an engine. The cylinder block consists of multiple cylinders, each with a piston inside. The pistons are arranged in a linear fashion, and the cylinders are lined up in a row. The image also shows a chain, which is likely part of the engine's mechanism or part of the engine's construction. The overall scene gives an impression of a detailed view of an engine's internal components.	8

Figure 3. Qualitative evaluation of different VLM on Automotive example 1 (motortrend, 2020)

User	Identify any signs of damage or wear in the brake components and explain their potential impact on vehicle performance.	GPT4 rating
LLaVA-1.5	In the image, there is a car wheel and a truck wheel with visible signs of wear, such as rust and dents. These signs of wear can have a negative impact on the vehicle's performance. Rust can weaken the metal components, making them more susceptible to breakage or failure. Dents can also affect the structural integrity of the wheel, potentially causing issues with the tire's ability to maintain proper contact with the road. Moreover, these signs of wear can lead to increased tire pressure, which may cause the tire to wear out faster or become less stable on the road. In extreme cases, a damaged wheel can lead to a complete tire failure, posing a significant safety risk to the vehicle's occupants and other road users. Regular maintenance and inspection of the brake components are crucial to ensure the safety and performance of the vehicle.	3
LLaVA-LLaMA	In the image, there is a close-up view of a brake component, which appears to be worn out. The worn-out brake component may have a damaged or worn-out pad, which can impact the vehicle's performance by reducing its ability to stop effectively. This can lead to longer stopping distances, increased risk of accidents, and potential damage to the vehicle's braking system. It is essential to replace the worn-out brake component with a new one to ensure safe and reliable vehicle operation.	5
LLaVA-Automotive	In the image, there are visible signs of rust and wear on the brake drum, which are components of the vehicle's brake system. Rust and wear can have a negative impact on vehicle performance, as they can cause the brake drum to become less effective at slowing down or stopping the vehicle. Rust can weaken the metal of the brake drum, making it more susceptible to cracking or warping, which can affect the drum's ability to absorb and dissipate heat generated during braking. Additionally, rust can cause the drum to become less responsive to the driver's input, leading to a longer stopping distance and potentially compromising the vehicle's overall braking performance. Wear on the brake drum can also affect the vehicle's braking performance. As the drum wears down, its surface becomes less effective at absorbing and dissipating heat, which can lead to a longer stopping distance and potentially compromised braking performance. Wear can also cause the drum to become less responsive to the driver's input, further affecting the vehicle's braking performance. To maintain optimal vehicle performance, it is important to regularly inspect and maintain the brake system, including cleaning and repairing any signs of rust or wear on the brake components.	6

Figure 4. Qualitative evaluation of different VLM on Automotive example 2 (thetimchannel, 2024)

sponses and the overall rationale behind them, based on its interpretation of the image. We found that LLaVA-Automotive performed better in qualitative evaluation in each one of them because of the domain-specific pretraining and instruction tuning. Two examples of the qualitative evaluation have been shown in Fig. 3 and Fig. 4.

5. CONCLUSION

This paper proposes Automotive-LLaVA, an instance of Diagnostics-LLaVA, a Visual language model capable of understanding Automotive domain images and providing suitable responses. We propose a methodology to create a domain-specific dataset by extracting images and corresponding texts from PDFs. We developed an automotive image-text pairs dataset for joint image and text representation by pretraining a CLIP-based ViT model. We also created an instruction dataset containing the images and corresponding question-answers towards instruction tuning of the VLM. We benchmarked an automotive dataset (ASE exam) for the evaluation of our models as well as the open-source model. We show that experimenting with the LLM used in an open source VLM can help gain good results in multi-modal question answering as depicted through quantitative evaluation on the benchmark. We also presented that the automotive multi-modal dataset constructed by us helped the model gain good understanding of domain-specific content, resulting in a good performance on qualitative evaluation.

6. LIMITATIONS AND FUTURE WORK

Our image-text pairs acquisition method requires some more improvements such that it can be generalized for extracting image-text pairs from all types of PDFs. As of now, it works with PDFs following a similar pattern of text and image that we observed in the books we used, but fails when an entirely different methodology of caption presence is detected, or if the images are embedded. Due to less amount of data captured during the data acquisition phase, we believe the model still has good scope for improvement by the addition of more unique data comprising of topics such as damage, repair, diagnosis, visual failure prediction, and more, from other knowledge sources.

We chose to pretrain and finetune the model from scratch by mixing our data with general multi-modal data openly available. There is a scope for training the existing model through LoRA as indicated by some recent work to reduce the computational burden, although the LoRA training would be restricted to the base LLM provided by the original authors of LLaVA.

It is possible that the addition of images in the original dataset with augmentation can hurt the overall model performance, therefore, more ablation study on the effect of data augmentation is required to determine optimum data augmentation of domain-specific images and text. We also believe that a large language model continued pretrained on a corpus of automotive text data can prove to be very valuable in producing domain-specific textual responses with higher accuracy. Therefore, we can conclude that using an automotive domain-specific LLM and automotive image dataset for VLM can help achieve better results.

REFERENCES

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., ... others (2023). Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... others (2020). Language models are few-shot learners. *Advances in neural information processing systems*, *33*, 1877–1901.

Dosovitskiy, A. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

freeasestudyguides. (2024). *Hub bearing ring.* Retrieved from `https://www.freeasestudyguides.com/a3-manual-transmission-test.html`

Guedes, G. B., & da Silva, A. E. A. (2021). Supervised learning approach for section title detection in pdf scientific articles. In *Advances in computational intelligence: 20th mexican international conference on artificial intelligence, micai 2021, mexico city, mexico, october 25–30, 2021, proceedings, part i 20* (pp. 44–54).

He, J., Wang, Y., Wang, L., Lu, H., He, J.-Y., Lan, J.-P., ... Xie, X. (2024). Multi-modal instruction tuned llms with fine-grained visual perception. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 13980–13990).

Kolo, E. (2006). *Does automotive service excellence (ase) certification enhance job performance of automotive service technicians?* (Unpublished doctoral dissertation). Virginia Polytechnic Institute and State University.

Kumar, A., & Starly, B. (2022). "fabner": information extraction from manufacturing process science domain literature using named entity recognition. *Journal of Intelligent Manufacturing*, *33*(8), 2393–2407.

Lai, Z., Bai, H., Zhang, H., Du, X., Shan, J., Yang, Y., ... Cao, M. (2024). Empowering unsupervised domain adaptation with large-scale pre-trained vision-language models. In *Proceedings of the ieee/cvf winter conference on applications of computer vision* (pp. 2691–2701).

Lee, J., Cha, S., Lee, Y., & Yang, C. (2024). Visual question answering instruction: Unlocking multimodal large language model to domain-specific visual multitasks. *arXiv preprint arXiv:2402.08360*.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... others (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, *33*, 9459–9474.

Liu, H., Li, C., Wu, Q., & Lee, Y. J. (2024). Visual instruction tuning. *Advances in neural information processing systems*, *36*.

Medeiros, T., Medeiros, M., Azevedo, M., Silva, M., Silva, I., & Costa, D. G. (2023). Analysis of language-model-powered chatbots for query resolution in pdf-based automotive manuals. *Vehicles*, *5*(4), 1384–1399.

MIT. (2024). *Understanding the visual knowledge of language models.* `https://news.mit.edu/2024/understanding-visual-knowledge-language-models-0617/`. ([Online; accessed 19-June-2024])

motortrend. (2020). *Engine cylinder block.* Retrieved from `https://www.motortrend.com/uploads/sites/21/2020/03/002-Difference-between-long-short-block.jpg`

Park, S.-M., & Kim, Y.-G. (2023). Visual language integration: A survey and open challenges. *Computer Science Review*, *48*, 100548.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... others (2021). Learning transferable visual models from natural language supervision. In *International conference on machine learning* (pp. 8748–8763).

Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. *Journal of big data*, *6*(1), 1–48.

thetimchannel. (2024). *Brake worn.* Retrieved from `https://openverse.org/image/958dcf66-f298-4413-85a7-957cf8474742`

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., ... others (2023). Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Vidyaratne, L., Lee, X. Y., Kumar, A., Watanabe, T., Farahat, A., & Gupta, C. (2024). Generating troubleshooting trees for industrial equipment using large language models (llm). In *2024 ieee international conference on prognostics and health management (icphm)* (pp. 116–125).

Wang, J., Liu, Z., Zhao, L., Wu, Z., Ma, C., Yu, S., ... others (2023). Review of large vision models and visual prompt engineering. *Meta-Radiology*, 100047.

Ye, Q., Xu, H., Xu, G., Ye, J., Yan, M., Zhou, Y., ... others (2023). 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., ... Huang, F. (2024). mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 13040–13051).

Yemaneab, T. (1997). *Employers' perceptions of automotive service excellence (ase) certification benefits*. University of Minnesota.

Zhang, J., Huang, J., Jin, S., & Lu, S. (2024). Vision-language models for vision tasks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Zhang, Y., Pan, J., Zhou, Y., Pan, R., & Chai, J. (2023). Grounding visual illusions in language: Do vision-language models perceive illusions like humans? In *Proceedings of the 2023 conference on empirical methods in natural language processing* (pp. 5718–5728).

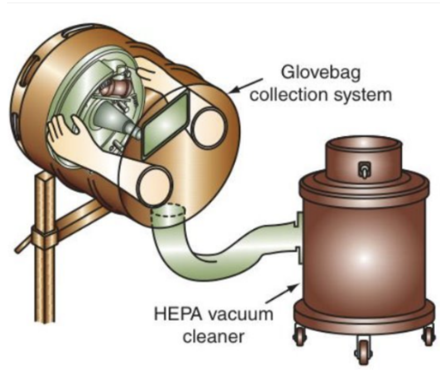
Zhang, Y., Zhang, R., Gu, J., Zhou, Y., Lipka, N., Yang, D., & Sun, T. (2023). Llavar: Enhanced visual instruction tuning for text-rich image understanding. *arXiv preprint arXiv:2306.17107*.

Zhao, R., Chen, H., Wang, W., Jiao, F., Do, X. L., Qin, C., ... others (2023). Retrieving multimodal information for augmented generation: A survey. *arXiv preprint arXiv:2303.10868*.

Zhao, X., Li, X., Duan, H., Huang, H., Li, Y., Chen, K., & Yang, H. (2024). *Mg-llava: Towards multi-granularity visual instruction tuning.*

APPENDIX

We use a set of instruction formats in the multimodal dataset for pretraining and finetuning of the Automotive-LLaVA model. The example instructions are shown in Fig. 5 for a special brake vacuum enclosure.



Mode	Pretraining conversation instruction
human	Describe the image concisely.
gpt	A special brake vacuum enclosure is used to remove and trap brake dust
human	Share an informative description of the image.
gpt	In the course of inspecting a vehicle's brake system, especially when examining wheel brake assemblies, it is crucial to use a specialized vacuum system with HEPA filters to capture and contain brake dust, ensuring it does not become airborne. This approach is part of a thorough inspection process that includes examining the brake linings, wheel cylinders, springs, and other components, as well as additional steps for larger vehicles such as trucks and vans with heavy-duty rear differentials.
human	Provide a detailed description of the image.
gpt	Whenever a vehicle is inspected for a brake system complaint, it is essential to check all wheel brake assemblies. This process involves removing the brake drums to thoroughly examine the brake linings, wheel cylinders, springs, and other components. If the customer's concern is unrelated to the drum brakes, like noise or other issues hinting at problems with the disc brakes or hydraulic system, it's not necessary to inspect the drum. An important aspect to consider while performing these checks is the collection of brake dust. It's highly recommended to use an approved method such as employing a vacuum system. These systems are equipped with high-efficiency particulate air (HEPA) filters, designed to trap brake dust and prevent it from becoming airborne. Working on larger vehicles such as pickup trucks and vans with one-ton heavy-duty rear differentials requires additional steps. To remove the drum in these vehicles, the rear axle shaft must be unbolted from the axle housing, and the wheel bearings removed. This process typically involves removing the axle shaft to drum bolts, pulling the axle from the differential, and setting it aside. The next step is to remove the retaining nut holding the bearings and drum to the axle tube, which might require a special socket to accomplish. Bear in mind that the drum in these larger vehicles is significantly heavier than those in average passenger cars. It's crucial to familiarize yourself with the procedure of removing and installing the drum before attempting to do so. Lastly, conducting a visual inspection of external components is equally important. You should examine the outside of the brake drum for any damage to the drum or cooling fins. Always refer to the service information for procedures specific to the vehicle to ensure a thorough and safe inspection.
human	Give a brief summary of the image.
gpt	A special brake vacuum enclosure is deployed to effectively remove and trap brake dust during the inspection and servicing of vehicle brake systems."

Mode	Finetuning conversation instruction
human	What is the illustrated device called?
gpt	The illustrated device is called a special brake vacuum enclosure.
human	What is the purpose of the HEPA filter in the brake vacuum system?
gpt	The HEPA filter in the brake vacuum system is designed to trap brake dust and prevent it from becoming airborne.
human	When inspecting a vehicle's brake system, when is it not necessary to inspect the drum brakes?
gpt	It is not necessary to inspect the drum brakes if the customer's concern is unrelated to them, like noise or other issues hinting at problems with the disc brakes or hydraulic system.
human	What additional steps are required when working on larger vehicles like pickup trucks and vans with heavy-duty rear differentials?
gpt	Additional steps include unbolting the rear axle shaft from the axle housing and removing the wheel bearings to remove the drum.
human	What should be examined during the visual inspection of external brake components?
gpt	During the visual inspection of external brake components, the outside of the brake drum should be examined for any damage to the drum or cooling fins.

Figure 5. Pretraining (Top) and Finetuning (Bottom) instruction prompt example for an image showing a special brake vacuum enclosure.