mPLUG-Owl🦉: Modularization Empowers Large Language Models with Multimodality

Qinghao Ye[*], Haiyang Xu[*], Guohai Xu[*], Jiabo Ye, Ming Yan[†], Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, Chenliang Li, Yuanhong Xu, Hehong Chen, Junfeng Tian, Qian Qi, Ji Zhang, Fei Huang
DAMO Academy, Alibaba Group
{yeqinghao.yqh, shuofeng.xhy, guohai.xgh, ym119608}@alibaba-inc.com

Abstract

Large language models (LLMs) have demonstrated impressive zero-shot abilities on a variety of open-ended tasks, while recent research has also explored the use of LLMs for multi-modal generation. In this study, we introduce **mPLUG-Owl**, a novel training paradigm that equips LLMs with multi-modal abilities through modularized learning of foundation LLM, a visual knowledge module, and a visual abstractor module. This approach can support multiple modalities and facilitate diverse unimodal and multimodal abilities through modality collaboration. The training paradigm of mPLUG-Owl involves a two-stage method for aligning image and text, which learns visual knowledge with the assistance of LLM, while maintaining and even improving the generation abilities of LLM. In the first stage, the visual knowledge module and abstractor module are trained with frozen LLM module to align the image and text. In the second stage, language-only and multi-modal supervised datasets are used to jointly fine-tune a low-rank adaption (LoRA) module on LLM and the abstractor module by freezing the visual knowledge module. We carefully build a visually-related instruction evaluation set **OwlEval**. Experimental results show that our model outperform existing multi-modal models, demonstrating mPLUG-Owl's impressive instruction and visual understanding ability, multi-turn conversation ability and knowledge reasoning ability. Besides, we observe some unexpected and exciting abilities such as multi-image correlation and scene text understanding, which makes it possible to leverage it for harder real scenarios, such as vision-only document comprehension. Our code, pre-trained model, instruction-tuned models, and evaluation set are available at https://github.com/X-PLUG/mPLUG-Owl. The online demo is available at https://www.modelscope.cn/studios/damo/mPLUG-Owl.

1 Introduction

Large language models (LLMs) such as GPT-3 [Brown et al., 2020], BLOOM [Scao et al., 2022], LLaMA [Touvron et al., 2023] have experienced rapid development to make general artificial intelligence possible, which demonstrates impressive zero-shot abilities on various linguistic applications. However, except GPT-4 [OpenAI, 2023], current general LLMs cannot support different modalities of input and develop impressive multimodal abilities.

Although GPT-4 [OpenAI, 2023] has exhibited remarkable multimodal abilities, the methods behind its extraordinary abilities remain a mystery. Recently, researchers have been extending LLMs

[*]Equal contribution
[†]Corresponding author

Preprint. Under review.

to understand visual inputs in two different paradigms: systematic collaboration and end-to-end trained models. However, systematic collaboration approaches, including Visual ChatGPT [Wu et al., 2023], MM-REACT [Yang et al., 2023], and HuggingGPT [Shen et al., 2023], are designed to facilitate the coordination of various vision models or tools to express visual information with text descriptions. However, these approaches may not be able to comprehend specific multimodal instructions due to their lack of alignment with different modalities. Additionally, these approaches may encounter challenges related to inference efficiency and cost. End-to-end models, such as BLIP-2 [Li et al., 2023], LLaVA [Liu et al., 2023], and MiniGPT-4 [Zhu et al., 2023a] aim to use unified models to support different modalities. However, these models have some limitations as they take frozen visual models, which may lead to inadequate alignment due to the limited number of parameters. Moreover, they cannot unlock various abilities due to missing unimodal and multimodal instruction.

In this paper, we present mPLUG-Owl with an innovative modularized training paradigm for large multi-modal language models that can support multiple modalities concurrently, drawing inspiration from the concept of modularization [Xu et al., 2023b, Li et al., 2022, Xu et al., 2021, Ye et al., 2022]. Our method harnesses the power of pre-trained LLM, visual knowledge module, and connected visual abstractor module to achieve effective alignment between images and text, and utilizes a two-stage training scheme to stimulate impressive unimodal and multimodal abilities. Our approach even enhances the strong generation abilities of LLM by modality collaboration between modalities. In the first step, we align the image and text to acquire comprehensive visual knowledge using text-image pairs, which is accomplished by training the visual knowledge module and abstractor module with the frozen LLM module. Subsequently, we fine-tune mPLUG-Owl with language-only and multi-modal instructions to unlock a range of unimodal and multimodal abilities. We freeze the visual knowledge module and train low-rank adaption (LoRA) [Hu et al., 2022] on LLM and visual abstractor module jointly. This approach allows for the effective integration of textual and visual information, facilitating the development of versatile and robust cognitive abilities.

Our experiments on a carefully-built visually related instruction evaluation set OwlEval shows that mPLUG-Owl outperforms existing models such as MiniGPT-4 [Zhu et al., 2023a] and LLaVA [Liu et al., 2023]. We separately verifies mPLUG-Owl's remarkable abilities in instruction understanding, visual understanding, knowledge transfer, and multi-turn dialogue. Abundant ablation study is performed to show the effectiveness of our training paradigm. Furthermore, we find some unexpected emerging ability such as multi-image correlation, multilingual conversation and scene text understanding.

Our main contributions can be highlighted as follows:

- We propose mPLUG-Owl, a novel training paradigm for large language models through modularization.
- We carefully construct an instruction evaluation set, dubbed **OwlEval**, to assess the capabilities of different models in the context of visual-related tasks.
- Experimental results demonstrate that mPLUG-Owl excels in multi-modal instruction understanding and multi-turn dialogue, surpassing the performance of existing models.

2 Related Work

2.1 Large Language Models

In recent times, Large Language Models (LLMs) have garnered increasing attention for their exceptional performance in diverse natural language processing (NLP) tasks. Initially, transformer models such as BERT [Devlin et al., 2019], GPT [Radford and Narasimhan, 2018], and T5 [Raffel et al., 2020] were developed with different pre-training objectives. However, the emergence of GPT-3 [Brown et al., 2020], which scales up the number of model parameters and data size, showcases significant zero-shot generalization abilities, enabling them to perform commendably on previously unseen tasks. Consequently, numerous LLMs such as OPT [Zhang et al., 2022], BLOOM [Scao et al., 2022], PaLM [Chowdhery et al., 2022], and LLaMA [Touvron et al., 2023] are created, ushering in the success of LLMs. Additionally, Ouyang et al. [Ouyang et al., 2022] propose InstructGPT by aligning human instruction and feedback with GPT-3. Furthermore, it has been applied to Chat-

GPT [OpenAI, 2022], which facilitates conversational interaction with humans by responding to a broad range of diverse and intricate queries and instructions.

2.2 Multi-Modal Large Language Models

Despite the successful applications of LLMs in natural language processing, it is still struggling for LLMs to perceive other modalities such as vision and audio. Recently, researchers have been extending language models to understand visual inputs in two different paradigms: systematic collaboration and end-to-end trained models. Systematic collaboration approaches, such as Visual ChatGPT [Wu et al., 2023], MM-REACT [Yang et al., 2023], and HuggingGPT [Shen et al., 2023], leverage various vision experts or tools to express visual information with text descriptions. Subsequently, large language models, such as ChatGPT, can act as the agents, and be prompted to select the appropriate experts and tools for visual understanding. Finally, LLMs would summarize the output of these experts to answer user queries. On the other hand, some approaches [Li et al., 2023, Alayrac et al., 2022, Liu et al., 2023] leverage the pre-trained large language model to build unified models for multi-modality. For example, Flamingo [Alayrac et al., 2022] freezes the pre-trained vision encoder and large language model and fuses vision and language modalities with gated cross-attention showing impressive few-shot capabilities. Additionally, BLIP-2 [Li et al., 2023] designs Q-Former to align the visual features from the frozen visual encoder and large language models with Flan-T5 [Chung et al., 2022] and OPT [Zhang et al., 2022]. Moreover, PaLM-E [Driess et al., 2023] directly inputs features from sensor modalities with PaLM [Chowdhery et al., 2022], which has 520 billion parameters, contributing to robust performance in real-world perceptions. Furthermore, some powerful instruction-tuned language models that built upon open-sourced foundation model LLaMA [Touvron et al., 2023], such as Alpaca [Taori et al., 2023] and Vicuna [Vicuna, 2023], exhibit comparable performance to ChatGPT [OpenAI, 2022] and GPT-4 [OpenAI, 2023]. MiniGPT-4 [Zhu et al., 2023a] and LLaVA [Liu et al., 2023] align these finetuned models with extracted visual features from the frozen visual backbone. In contrast, mPLUG-Owl not only aligns the representation between the vision and language foundation model (e.g. CLIP and LLaMA) in terms of knowledge acquisition and grounding to the real world but also can understand language and multi-modal instructions, showcasing strong zero-shot generalization and multi-turn conversation capabilities.

3 mPLUG-Owl

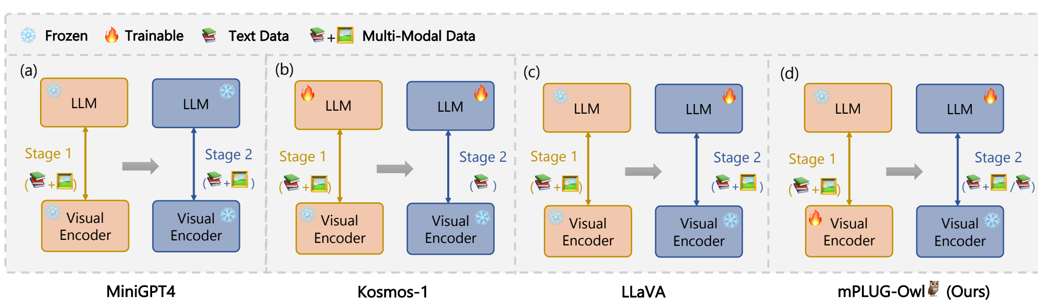


Figure 1: Comparison between different training paradigms. All of these methods are trained in a two-stage fashion. Stage 1 stands for pre-training and Stage 2 represents instruction tuning.

3.1 Architecture Overview

As illustrated in Figure 1, there exist mainly three types of end-to-end multimodal LLMs: 1) models that utilize limited parameters with frozen LLM and visual models during pretraining and instruction tuning, such as MiniGPT4; 2) models that incorporate trainable LLMs and frozen visual models, exemplified by Kosmos-1; and 3) models that involve trainable LLMs during instruction tuning and frozen visual models, as seen in LLaVA. Nevertheless, these models exhibit certain constraints since they depend on frozen visual models, which can lead to insufficient alignment due to the limited number of parameters. Furthermore, they fail to effectively stimulate a diverse set of abilities, as they lack both unimodal and multimodal instruction.

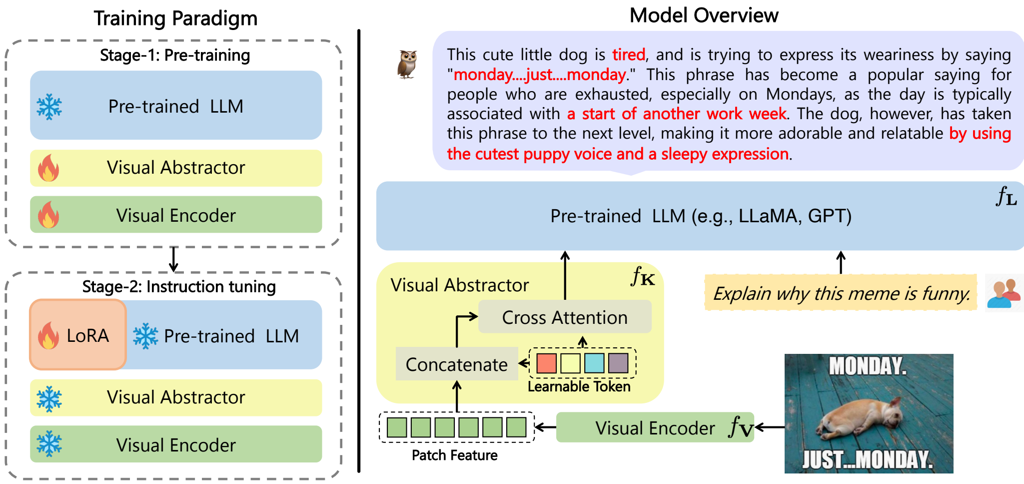


Figure 2: Our training paradigm and model overview.

To this end, we propose mPLUG-Owl, a multi-modal language model that is capable of perceiving various modalities while taking the visual context and information into account and generating corresponding outputs. Specifically, as illustrated in Figure 2, mPLUG-Owl consists of a vision foundation model $f_{\mathbf{V}}$ to encode the visual knowledge, a language foundation model $f_{\mathbf{L}}$, and a visual abstractor module $f_{\mathbf{K}}$. We first obtain dense image representations from the pre-trained visual foundation model $f_{\mathbf{V}}$. However, such dense features would fragment the fine-grained image information and bring large computation due to the lengthy sequence when feeding into $f_{\mathbf{L}}$. To mitigate this issue, we employ the visual abstractor module $f_{\mathbf{K}}$ to summarize visual information within several learnable tokens, thereby obtaining higher semantic visual representations and reducing computation, as illustrated in Figure 2. The visual representations are combined with text queries and fed into the language model to generate the response.

3.2 Training Scheme

Multimodal Pretraining Large-scale language models, such as GPT-3 [Brown et al., 2020] and LLaMA [Touvron et al., 2023], are trained on extensive and diverse data collected from the internet, providing them with a comprehensive understanding of the world. This vast knowledge base endows these models with remarkable capabilities across a range of tasks. However, the utilization of visual information in such models remains underexplored. Previous approaches [Zhu et al., 2023a, Liu et al., 2023] have employed a limited number of additional parameters to learn the alignment between visual data and language models, constraining their capacity to comprehend complex visual information. To enhance the ability of large-scale language models to perceive visual information while integrating their internal abilities, we propose a novel training paradigm that incorporates a trainable visual backbone $f_{\mathbf{V}}$ and an additional visual abstractor $f_{\mathbf{K}}$, while maintaining the pre-trained language model $f_{\mathbf{L}}$ in a frozen state. This approach enables the model to effectively capture both low-level and higher semantic visual information and align it with the pre-trained language model without compromising its performance.

Joint Instruction Tuning Upon completion of the prior phase, the model acquires the ability to retain a considerable amount of knowledge and provide reasonable answers to human queries. Nonetheless, it continues to exhibit challenges in generating coherent linguistic responses. As posited in GPT-3 [Brown et al., 2020], refining the model through instruction tuning is essential for accurately discerning user intentions. Previous attempts [Li et al., 2022, Xu et al., 2023b] in multi-modal learning have demonstrated that joint learning from uni-modal and multi-modal sources can lead to significant improvements owing to the collaboration between different modalities. Building on this insight, we present a novel vision-language joint instruction tuning strategy to facilitate better alignment between mPLUG-Owl and human instructions and intentions. Specifically, given that the model can comprehend the visual concepts and knowledge depicted in images through visual knowledge learning, we freeze the entire model and employ low-rank adaption (i.e., LoRA [Hu et al., 2022]) to adapt $f_{\mathbf{L}}$ by training multiple low-rank matrices for efficient alignment with human instructions. For each data record, we unified them in a snippet of conversation following Vicuna [Vicuna, 2023], and we compute the loss on the response. During the training, we accumulate the gradient for text-only instruction data and multi-modal instruction data for multiple batches and updated the parameters. Therefore, by joint training with both language and multi-modal instructions, mPLUG-Owl can better understand a wide range of instructions and respond with more natural and reliable output. Moreover, our approach can easily handle various text and multi-modal instructions without the need for realignment of the vision and language models, as required by methods such as MiniGPT-4 [Zhu et al., 2023a] and LLaVA [Liu et al., 2023].

Training Objective The model is trained using the language modeling task, which entails learning to generate subsequent tokens based on the preceding context. The primary objective of the training process is to maximize the log-likelihood of the tokens. It is important to note that only discrete tokens, such as text tokens, are considered in the calculation of the training loss. Most significantly, the emergence of diverse capabilities resulting from the training task during the joint instruction tuning stage enhances the performance of mPLUG-Owl in downstream applications.

4 Experiment

4.1 Experimental Setup

Model Settings. We choose ViT-L/14 [Dosovitskiy et al., 2021] as the visual foundation model $f_{\mathbf{V}}$ which has 24 layers with hidden dimension set as 1024 and patch size set as 14. For faster convergence, the ViT is initialized from CLIP ViT-L/14 model pre-trained via contrastive learning. Different with LLaVA [Liu et al., 2023] and MiniGPT-4 [Zhu et al., 2023a], to demonstrate the effectiveness and generalization ability, we utilize raw LLaMA-7B [Touvron et al., 2023] rather than its instruction-tuned variants such as Alpaca [Taori et al., 2023] and Vicuna [Vicuna, 2023]. The total number of parameters of mPLUG-Owl is about 7.2B. More details about hyper-parameters can be found in Appendix.

Data and Training Details. For the first stage, we utilize the image-caption pairs from several datasets, including LAION-400M [Schuhmann et al., 2021], COYO-700M [Byeon et al., 2022], Conceptual Captions [Sharma et al., 2018] and MSCOCO [Chen et al., 2015]. We use a batch size of 2.1 million tokens and train mPLUG-Owl for 50k steps, corresponding to about 104 billion tokens. We adopt the AdamW optimizer with $\beta = (0.9, 0.98)$, and set the learning rate and weight decay to 0.0001 and 0.1 respectively. We warm up the training with 2k warm-up steps then decay the learning rate with the cosine schedule. The input image is randomly resized to 224×224. Besides, we tokenize the text input with SentencePiece [Kudo and Richardson, 2018] tokenizer. For the second stage, we gather pure text instruction data from three distinct sources: 102k data from the Alpaca [Taori et al., 2023], 90k from the Vicuna [Vicuna, 2023], and 50k from the Baize [Xu et al., 2023a]. Additionally, we utilize 150k multi-modal instruction data from the LLaVA dataset [Liu et al., 2023]. We train mPLUG-Owl for 2k steps with the batch size 256, and the learning rate is set to 0.00002.

Baselines. We compare our mPLUG-Owl with end-to-end models and systematic collaboration approaches as follows:

- *OpenFlamingo* [Zhu et al., 2023b] is an open-source version of Flamingo [Alayrac et al., 2022] model. We use the released code of OpenFlamingo-9B[3] to run zero-shot generation.
- *BLIP-2* [Li et al., 2023] is pre-trained through bootstrapped learning from off-the-shelf frozen pre-trained image models and large language models using an efficient pre-training strategy. We use the released code of BLIP-2 ViT-G FlanT5$_{XXL}$[4] to perform zero-shot generation.
- *MiniGPT-4* [Zhu et al., 2023a] utilizes a single projection layer to align visual information from a pre-trained vision encoder with LLM. Specifically, they employ the same visual

[3] https://github.com/mlfoundations/open_flamingo

[4] https://github.com/salesforce/LAVIS/tree/main/projects/blip2

encoder as used in BLIP-2, a ViT coupled with their pre-trained Q-Former, and Vicuna as LLM. We use the released demonstration[5] to perform image-instruction generation.

- *LLaVA* [Liu et al., 2023] applies a single projection layer to convert image features from pre-trained CLIP visual encoder ViT-L/14 into the language embedding space of Vicuna. We use their released demonstration[6] to perform image-instruction generation.
- *MM-REACT* [Yang et al., 2023] integrates ChatGPT/GPT-4 with various specialized vision experts to achieve multimodal reasoning and action. We use their released demonstration[7] to get responses.

4.2 Quantitative analysis

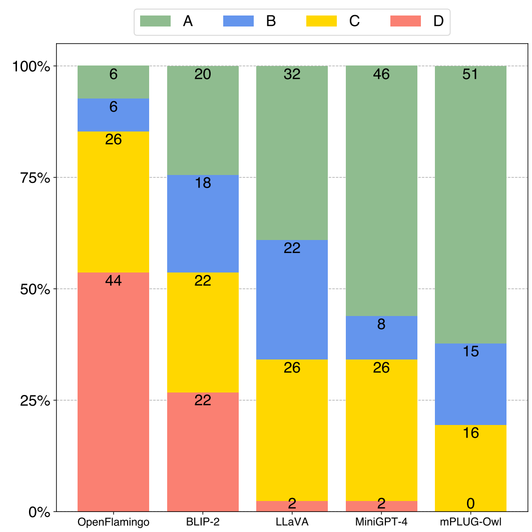


Figure 3: The comparison between mPLUG-Owl and baselines on OwlEval with manual evaluation metrics. The order of response quality ranking is as follows: A > B > C > D.

In order to comprehensively evaluate various models, we construct a visually-related evaluation set **OwlEval** by collecting 82 artificially constructed questions based on 50 images, where 21 from MiniGPT-4, 13 from MM-REACT, 9 from BLIP-2, 3 from GPT-4 and 4 collected by us. Partial images have multiple rounds of questions, refers to multi-turn conversation cases. These questions examine a variety of model capabilities including natural image understanding, diagram and flowchart comprehension, optical character recognition (OCR), multi-modal creation, knowledge-intensive QA, and referential interaction QA. As questions are open-ended, we employ manual evaluation metrics to rate the model's responses as A, B, C, or D following the rating method proposed in Self-Instruct [Wang et al., 2022].

We manually score 82 responses given by mPLUG-Owl and baselines. The comparison results are shown in Figure 3. First, mPLUG-Owl gets 66 A and B, while the most competitive baseline MiniGPT-4 gets 54. Second, mPLUG-Owl doesn't get any D scores, outperforming all the models. These results suggest that mPLUG-Owl can better understand both instructions and images, which results in a stronger capability in generating satisfactory responses. For a fair comparison, we have excluded those cases in which MM-REACT failed to make predictions. The results are shown separately in Figure 15 and mPLUG-Owl still exhibits superior performance.

To separately examine the single-turn and multi-turn conversation capabilities, we reorganize 82 questions into a single-turn conversation set and a multi-turn conversation set. The former contains the first question from 50 images. The latter contains 52 questions from multi-turn conversation cases. As shown in Figure 4, the mPLUG-Owl achieves outstanding performance in both single-turn and multi-turn conversations.

[5] https://huggingface.co/spaces/Vision-CAIR/minigpt4
[6] https://llava.hliu.cc
[7] https://huggingface.co/spaces/microsoft-cognitive-service/mm-react

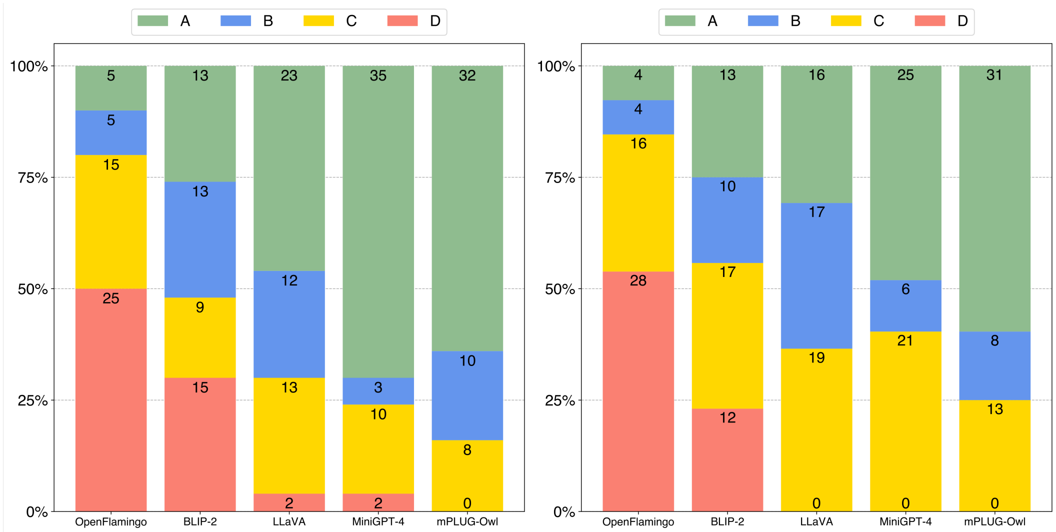


Figure 4: The comparison results of 50 single-turn responses (left) and 52 multi-turn responses (right) among mPLUG-Owl and baselines on OwlEval with manual evaluation metrics.

4.3 Ablation Study

We ablate the two-stage training scheme and the data modality of instruction tuning. Six dimensions of abilities are defined to complete visually related tasks, as shown in Table 1. For each question, we manually label the required abilities and annotate which abilities are reflected in the model's response. Table 2 shows the ability accuracy of different variants of mPLUG-Owl.

Training Strategy Ablation. As shown in Table 2, without joint instruction tuning, the model is not good at instruction understanding and fail to generalize pre-training abilities to other tasks (r1 vs r5). With the instruction tuning alone, although the model can better comprehend instructions, the model is incapable of achieving promising performance in visual knowledge-related tasks due to lacking of visually-related knowledge pretraining (r2 vs r5). With both multimodal pretraining and joint instruction tuning, the model achieves the best performance and demonstrates the effectiveness of our two-stage training scheme.

Instruction Data Ablation. By comparing r3 with r4, text-only instruction tuning brings more improvement in instruction understanding, while multi-modal instruction tuning achieves better knowledge and reasoning capabilities. This is due to that visual question answering mainly requires the alignment of vision and language knowledge, which is not optimized during text-only instruction tuning. Besides, we also verify that introducing multi-modal data during instruction tuning could further improve the model's performance on text-only tasks, as shown in Table 3 (r5 vs r4). Concretely, following the evaluation setting as Vicuna[Vicuna, 2023], for each question, we pair the response of each model with the one given by ChatGPT and prompt ChatGPT[8] to give two scores respectively for these two responses. Table 3 shows the total score and the score ratio with the ChatGPT score as a reference.

4.4 Qualitative Analysis

In this section, we show qualitative results from our evaluation set OwlEval.

Knowledge-intensive QA As shown in Figure 5, the instruction expects the model to identify the movie characters in the image. MM-REACT is unable to provide an effective response to the instruction, while MiniGPT-4 understands the instruction but failed to answer the movie characters. In contrast, mPLUG-Owl answers four out of the five characters present in the image. This demonstrates that mPLUG-Owl has a better understanding of the knowledge in the image.

[8] Without access to the GPT-4, we use the ChatGPT as the suboptimal scorer.

	Meaning	Definition
IU	Instruction Understanding	1. Understand text instruction. 2. Do not require a correct answer, but the response should be related to the instruction.
VU	Visual Understanding	1. Identify image information. 2. The answer faithfully reflects over 60% of visual information in the image.
OCR	Optical Character Recognition	1. Recognize text information in the image. 2. The answer faithfully reflects over 60% of text information in the image.
KTA	Knowledge Transfer Ability	1. Transfer knowledge between language and vision. (1) understand textual and visual content (2) align and transfer visual and language knowledge 2. Answers are mostly accurate with accuracy rate over 80%.
RA	Reasoning Ability	1. Combine image and text for reasoning. (1) understand textual and visual content (2) conduct multi-step reasoning (3) generate answers based on multi-step reasoning process 2. The final answer is essentially correct, but it lacks an explicit reasoning process. Alternatively, the final answer is mostly correct, and the reasoning process is over 80% accurate. 3. For example (1) Commonsense Knowledge Reasoning (2) Counterfactual Reasoning (3) Spatial Relation Reasoning (4) Numerical Computation (5) Coding
MDA	Multi-turn Dialogue Ability	1. Understand instructions and handle multi-turn conversations. 2. It includes clear references to multiple conversations and handles natural language semantics in context effectively. 3. The semantics and references are mostly correct, with an accuracy rate of over 80%.

Table 1: The definition of 6 abilities to complete visually-related tasks.

	Multimodal Pretraining	**Pure Text Instruction**	**Multi-modal Instruction**	**Ability**					
				IU	**VU**	**OCR**	**KTA**	**RA**	**MDA**
r1	✓			$58.5_{(-41.5)}$	$38.1_{(-57.1)}$	$13.3_{(-43.4)}$	$16.7_{(-70.8)}$	$17.1_{(-62.9)}$	$40.0_{(-55.0)}$
r2		✓	✓	$93.9_{(-6.1)}$	$47.6_{(-47.6)}$	$23.3_{(-33.4)}$	$29.2_{(-58.3)}$	$14.3_{(-65.7)}$	$45.0_{(-50.0)}$
r3	✓	✓		$93.0_{(-7.0)}$	$73.0_{(-22.2)}$	$40.0_{(-16.7)}$	$41.7_{(-45.8)}$	$48.6_{(-31.4)}$	$80.0_{(-15.0)}$
r4	✓		✓	$86.6_{(-13.4)}$	$68.3_{(-26.9)}$	$40.0_{(-16.7)}$	$50.0_{(-37.5)}$	$60.0_{(-20.0)}$	$75.0_{(-20.0)}$
r5	✓	✓	✓	**100.0**	**95.2**	**56.7**	**87.5**	**80.0**	**95.0**
MiniGPT-4 [Zhu et al., 2023a]				97.6	81.0	40.0	83.3	65.7	75.0

Table 2: The ablation results. Each value represents the proportion of questions where the corresponding ability is correctly reflected in the model's response. IU: instruction understanding, VU: visual understanding, OCR: optical character recognition, KTA: knowledge transferability, RA: reasoning ability, MDA: multi-turn dialogue ability.

Multi-round Conversation The instruction in Figure 6 requires the model to identify the content of the image based on the referential information. The baseline models often made mistakes when faced with referential expressions related to spatial orientation, human behavior, and target attributes in the questions, whereas mPLUG-Owl provided the most accurate response. This capability stems from mPLUG-Owl's fine-grained understanding of the image, allowing it to locate the corresponding part of the image based on the referential information in the instruction.

Reasoning Figure 7 shows an instruction asking models to give a prediction based on visual information and explain the reason. mPLUG-Owl analyzes the characteristics of the two teams from

Model	Tuning Strategy	Model Score	ChatGPT Score	Ratio
Alpaca-7B	Full	573	708	80.93%
Vicuna-7B	Full	612	684	89.47%
mPLUG-Owl w/o multimodal tuning (r4)	LoRA	587	682	86.07%
mPLUG-Owl (r5)	LoRA	600	692	86.71%

Table 3: The performance of 80 text-only questions from Vicuna[Vicuna, 2023] assessed by ChatGPT.

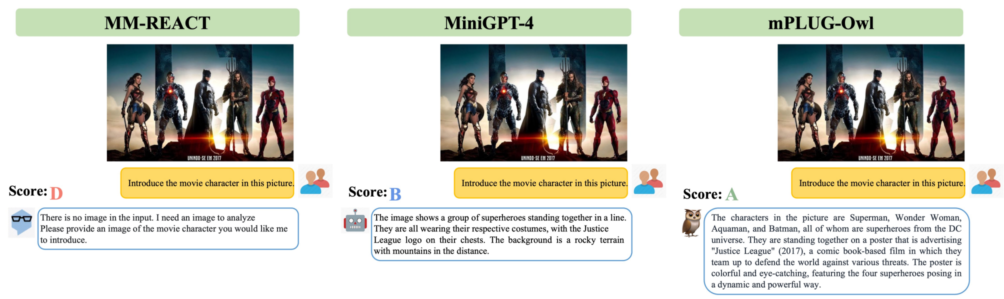


Figure 5: A comparison of Knowledge-intensive QA.

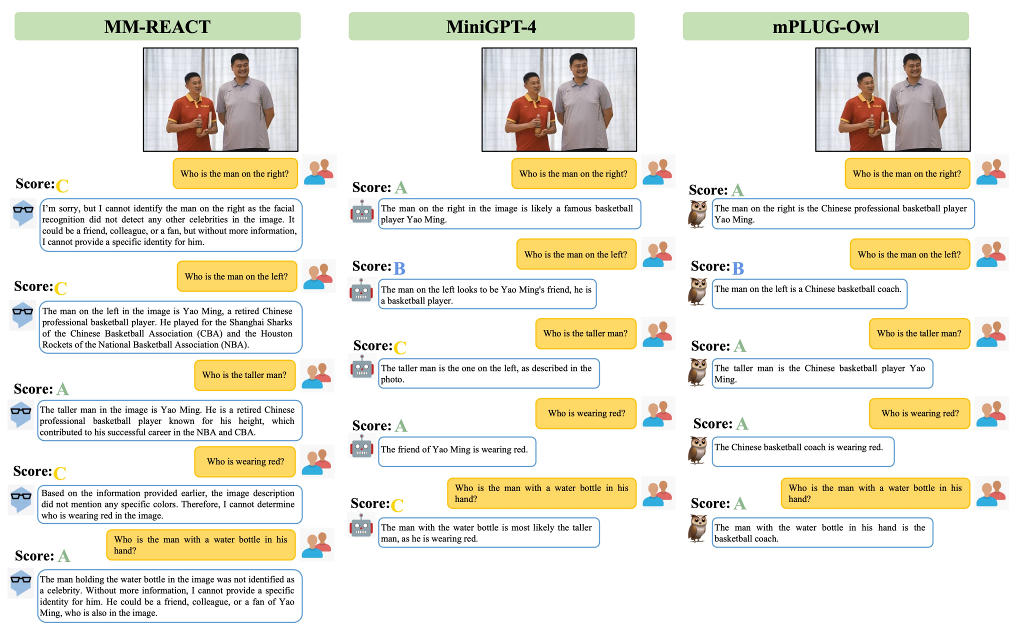


Figure 6: A comparison of Multi-turn Conversation.

the aspects of the lineup and tactics and uses them to reason for the outcome. Although MiniGPT-4 also performs well, its persuasiveness in reasoning is slightly inferior to mPLUG-Owl.

Joke Comprehension The case in Figure 8 comes from the GPT-4[OpenAI, 2023], which requires the model to understand and explain a visually related joke. GPT-4 not only follows the instructions in performing analysis panel by panel but also almost perfectly understands the humor of the charging method. mPLUG-Owl also understands this unusual humor, but it incorrectly identified the "VGA" to "USB". This is mainly due to the limitation of visual information in our training data. More cases about joke comprehension are shown in Figure 9.

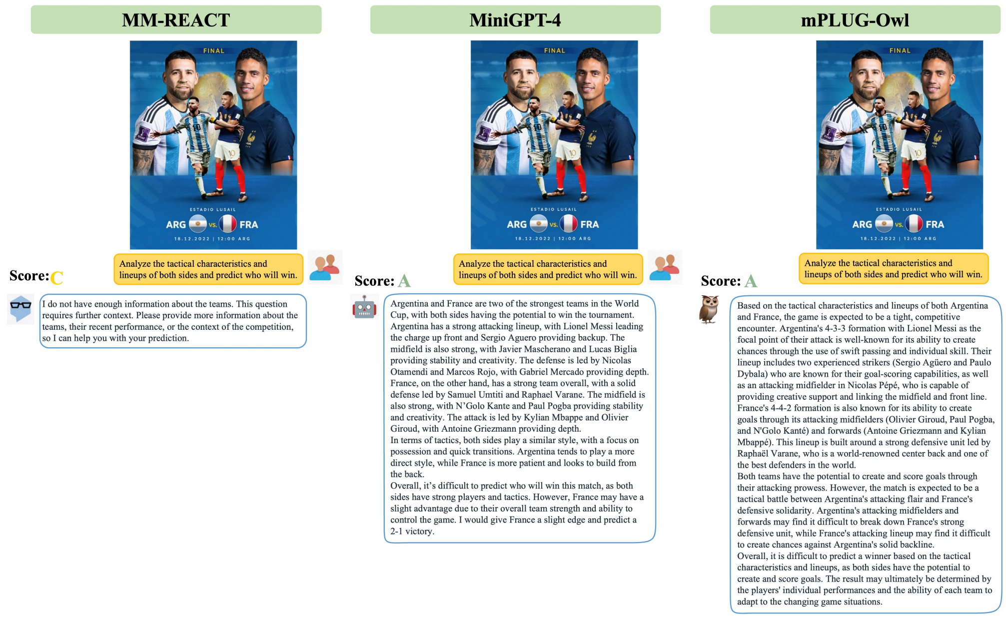


Figure 7: A comparison of Reasoning QA.

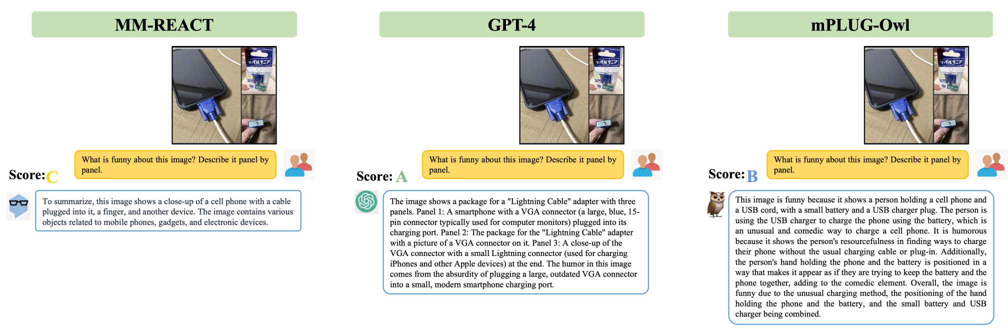


Figure 8: A comparison of Joke Understanding.

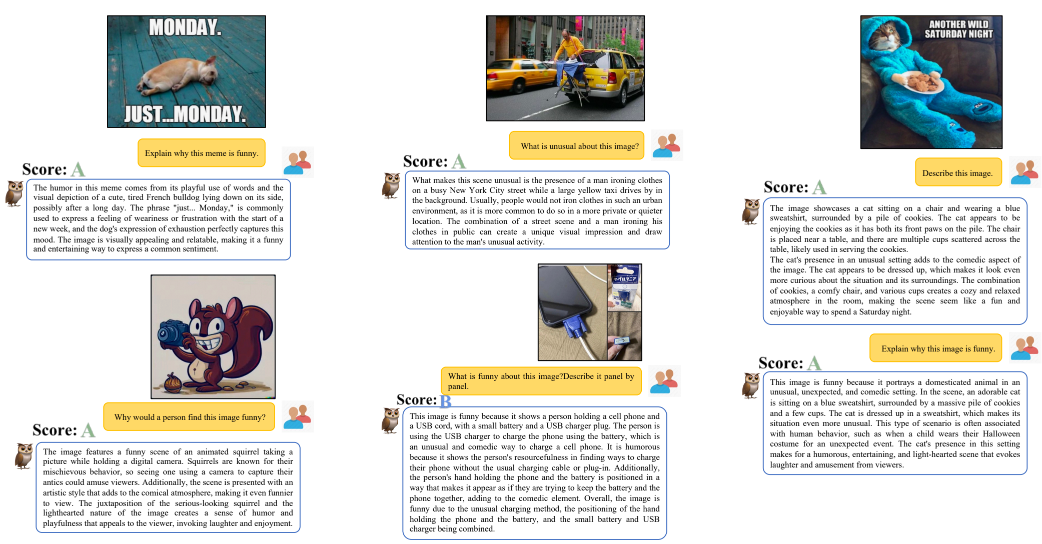


Figure 9: More cases of Jokes Comprehension by mPLUG-Owl.

5 Discussion and Limitation

In this section, we show some nascent abilities of mPLUG-Owl that is not yet fully developed and discuss the limitation. Part of cases (without scores) in this section are not in OwlEval.

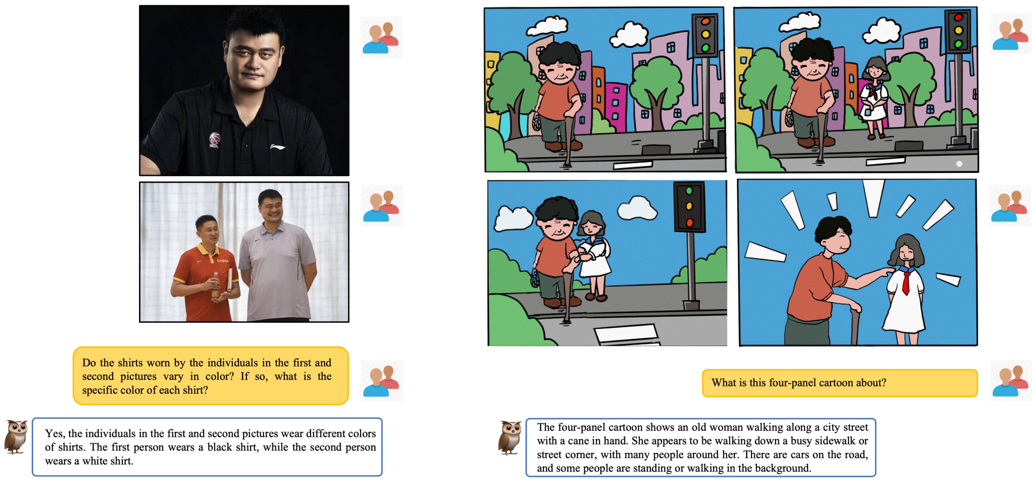


Figure 10: Multi-image correlation cases.

Multi-image Correlation In Figure 10, mPLUG-Owl shows a emerging but not strong vision correlation capability across multiple images. In the left case, the model could identify an identical person in two images and correctly tell the difference of cloth color. But in the left case, the model fails to relate 4 images and produces some text hallucinations.

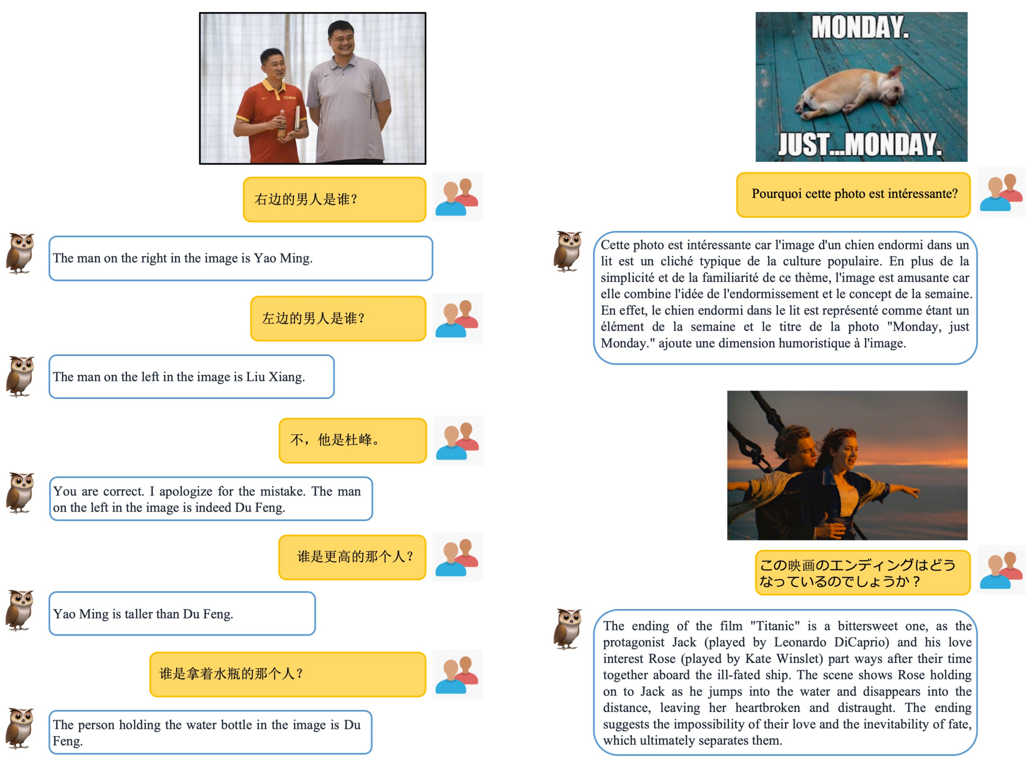


Figure 11: Example prompt of multilingual understanding which showcases the multilingual abilities across Chinese, French, and Japanese, respectively.

Multilingual Conversation Besides English, we further test the model's multilingual ability. As shown in Figure 11, although there is no multilingual data during our two-stage training, mPLUG-Owl shows a promising multilingual understanding for Chinese, French and Japanese. We mainly attribute this ability to the raw text knowledge in LLaMa[Touvron et al., 2023]. However, due to the lacking of multilingual training, mPLUG-Owl may fail to response in corresponding languages.

Scene Text Understanding In Figure 16, mPLUG-Owl demonstrates its OCR ability in some simple scenes, but we can see that the model's perception of numbers in images is still limited. However, for the OCR of complex scenes, as shown in Figure 17-18, the performance of mPLUG-Owl is more general, mainly because the perception of numbers in images is weak, which affects the subsequent reasoning calculation.

Vision-only Document Comprehension Although we did not use any document annotation data for training, the model exhibited some text recognition and document understanding capabilities. Hence, we delved deeper into the combination of document understanding and functionality of our model. as illustrated in Figure 12, we explored movie review writing, code generation, code explanation, chat summary, and application guidance. The model show decent performance in (a) and (b), but still, had some errors. Meanwhile, it was unable to provide usable responses in (d), (e), and (f). Therefore, there is further scope to explore our model's potential in document understanding and downstream applications.

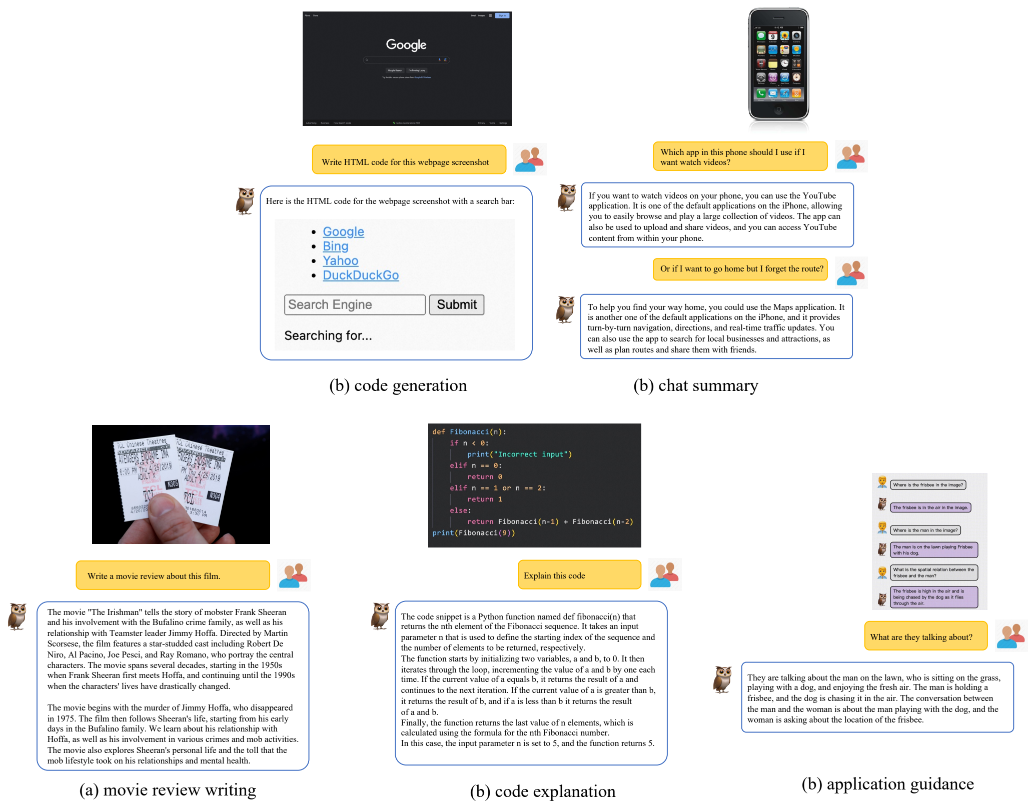


Figure 12: Examples about various document understanding and application.

Open-ended Creation mPLUG-Owl performs well in the creation of poetry, lyrics, advertisements and other works based on images. Its performance in some cases is shown in Figure 13-14. However, further exploration is needed for more functional and practical creations.

Figure 13: Open-ended creation cases.

6 Conclusion

We propose mPLUG-Owl, a novel training paradigm that enhances the multi-modal abilities of large language models (LLMs). Our approach consists of modularized learning of foundation LLM, a visual knowledge module, and a visual abstractor module, which can support multiple modalities and facilitate diverse unimodal and multimodal abilities through modality collaboration. We employ a two-stage method for aligning image and text, which learns visual knowledge with the assistance of LLM while maintaining and even improving the generation abilities of LLM. Experimental results demonstrate the impressive capabilities of mPLUG-Owl, indicating its potential for various applications in multi-modal generation.

Write a guide to travel here.

The Potala Palace in Tibet is an architectural masterpiece that stands as the most iconic and recognizable landmark in the region. Located in the Lhasa Valley, it is a UNESCO World Heritage Site and a symbol of Tibetan cultural heritage.
The Potala Palace was originally constructed as the winter residence of the Dalai Lama, the spiritual leader of Tibetan Buddhism, in the 7th century, and it has since served as a residence for the ruling dynasties of Tibet. The palace is made up of several buildings, including the White Palace, which was the residence of the Dalai Lama himself, and the Red Palace, which was the administrative center for the ruling government.
The Potala Palace is renowned for its elaborate architecture and artwork. The palace's exterior is covered with thousands of intricate carvings, murals and statues, and its interior is adorned with precious Buddhist relics, paintings, and murals.
The palace also boasts stunning views of the surrounding landscape, making it an ideal destination for visitors looking to take in the beauty of the Tibetan countryside and experience a piece of its unique history and culture.
To visit the Potala Palace, one must obtain a travel visa to Tibet, as the region is currently only open to foreign visitors on a limited basis. Once in Tibet, tourists will have to obtain a travel permit from the Tibetan Government to visit the palace.

Write a long tweet to recommend this movie to your friends.

I'd recommend Avatar: The Last Airbender for friends who enjoy action-packed movies with stunning visuals and a unique storyline. The movie takes place in a world of elemental magic and follows the protagonists' struggle to defend their people against the oppressive forces that seek to exploit their resources. The film's diverse cast, including the protagonist, who is a young Avatar with the ability to control all four elements, and a group of friends and allies who support him, make for a dynamic and entertaining watch. It's a great movie for fans of fantasy, adventure, and martial arts.

Los Angeles Times

latimes.com

Loyal Sessions joins Trump's list of targets

President's remarks against key officials reinforce the notion no one close to the Oval Office is safe.

McCain absence shrinks GOP's margin

Until his return, the Senate majority's agenda is imperiled.

O.J. Simpson is given parole after 9 years

Fallen star could be out of prison for Vegas robbery by October.

Write a newscast based on the picture.

The Los Angeles Times is reporting that O.J. Simpson, the infamous American former football player and actor, has been paroled from the Nevada State Prison after spending nearly nine years incarcerated for a 2007 armed robbery.
Simpson was granted parole on Thursday by Nevada's Parole Board, which cited his positive behavior and good conduct during his time in prison. Simpson's release comes as President Donald Trump recently announced a new list of potential pardons, which included Simpson.
In the wake of the news, Simpson's longtime friend, Al Sharpton, issued a statement expressing support for the parole decision, saying that Simpson has been "rehabilitated" and deserves a second chance in life. Sharpton added that Simpson's release is a victory for "all Americans who believe in redemption and second chances."
However, not everyone is celebrating Simpson's release. Many people, including some of the victims of his crimes, have voiced opposition to his release, saying that it undermines the rule of law and sends a negative message to society.
As a result of his parole, Simpson will have several restrictions placed on his freedom, including a curfew, a ban on alcohol consumption, and the requirement to wear an electronic monitoring bracelet. Simpson will also be required to stay in Nevada for the first year of his parole and will not be allowed to return to California, where the murders of Nicole Brown Simpson and Ron Goldman took place in 1994. Simpson will also be required to participate in a mental health program and complete community service work.
Overall, the release of O.J. Simpson has sparked controversy and divided public opinion. While some believe that Simpson has been rehabilitated and deserves a second chance, others believe that his release sends the wrong message about the importance of accountability for crimes committed.

Figure 14: Copywriting cases.

References

J. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds, R. Ring, E. Rutherford, S. Cabi, T. Han, Z. Gong, S. Samangooei, M. Monteiro, J. Menick, S. Borgeaud, A. Brock, A. Nematzadeh, S. Sharifzadeh, M. Binkowski, R. Barreira, O. Vinyals, A. Zisserman, and K. Simonyan. Flamingo: a visual language model for few-shot learning. *CoRR*, abs/2204.14198, 2022.

T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. Language models are few-shot learners. In *NeurIPS*, 2020.

M. Byeon, B. Park, H. Kim, S. Lee, W. Baek, and S. Kim. Coyo-700m: Image-text pair dataset. https://github.com/kakaobrain/coyo-dataset, 2022.

X. Chen, H. Fang, T. Lin, R. Vedantam, S. Gupta, P. Dollár, and C. L. Zitnick. Microsoft COCO captions: Data collection and evaluation server. *CoRR*, abs/1504.00325, 2015.

A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton, S. Gehrmann, P. Schuh, K. Shi, S. Tsvyashchenko, J. Maynez, A. Rao, P. Barnes, Y. Tay, N. Shazeer, V. Prabhakaran, E. Reif, N. Du, B. Hutchinson, R. Pope, J. Bradbury, J. Austin, M. Isard, G. Gur-Ari, P. Yin, T. Duke, A. Levskaya, S. Ghemawat, S. Dev, H. Michalewski, X. Garcia, V. Misra, K. Robinson, L. Fedus, D. Zhou, D. Ippolito, D. Luan, H. Lim, B. Zoph, A. Spiridonov, R. Sepassi, D. Dohan, S. Agrawal, M. Omernick, A. M. Dai, T. S. Pillai, M. Pellat, A. Lewkowycz, E. Moreira, R. Child, O. Polozov, K. Lee, Z. Zhou, X. Wang, B. Saeta, M. Diaz, O. Firat, M. Catasta, J. Wei, K. Meier-Hellstern, D. Eck, J. Dean, S. Petrov, and N. Fiedel. Palm: Scaling language modeling with pathways. *CoRR*, abs/2204.02311, 2022.

H. W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, E. Li, X. Wang, M. Dehghani, S. Brahma, A. Webson, S. S. Gu, Z. Dai, M. Suzgun, X. Chen, A. Chowdhery, S. Narang, G. Mishra, A. Yu, V. Y. Zhao, Y. Huang, A. M. Dai, H. Yu, S. Petrov, E. H. Chi, J. Dean, J. Devlin, A. Roberts, D. Zhou, Q. V. Le, and J. Wei. Scaling instruction-finetuned language models. *CoRR*, abs/2210.11416, 2022.

J. Devlin, M. Chang, K. Lee, and K. Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT (1)*, pages 4171–4186. Association for Computational Linguistics, 2019.

A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*. OpenReview.net, 2021.

D. Driess, F. Xia, M. S. M. Sajjadi, C. Lynch, A. Chowdhery, B. Ichter, A. Wahid, J. Tompson, Q. Vuong, T. Yu, W. Huang, Y. Chebotar, P. Sermanet, D. Duckworth, S. Levine, V. Vanhoucke, K. Hausman, M. Toussaint, K. Greff, A. Zeng, I. Mordatch, and P. Florence. Palm-e: An embodied multimodal language model. *CoRR*, abs/2303.03378, 2023.

E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen. Lora: Low-rank adaptation of large language models. In *ICLR*. OpenReview.net, 2022.

T. Kudo and J. Richardson. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *EMNLP (Demonstration)*, pages 66–71. Association for Computational Linguistics, 2018.

C. Li, H. Xu, J. Tian, W. Wang, M. Yan, B. Bi, J. Ye, H. Chen, G. Xu, Z. Cao, J. Zhang, S. Huang, F. Huang, J. Zhou, and L. Si. mplug: Effective and efficient vision-language learning by cross-modal skip-connections. In *EMNLP*, pages 7241–7259. Association for Computational Linguistics, 2022.

J. Li, D. Li, S. Savarese, and S. C. H. Hoi. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. *CoRR*, abs/2301.12597, 2023.

H. Liu, C. Li, Q. Wu, and Y. J. Lee. Visual instruction tuning. *CoRR*, abs/2304.08485, 2023.

OpenAI. Introducing chatgpt. https://openai.com/blog/chatgpt, 2022.

OpenAI. GPT-4 technical report. *CoRR*, abs/2303.08774, 2023.

L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. F. Christiano, J. Leike, and R. Lowe. Training language models to follow instructions with human feedback. *CoRR*, abs/2203.02155, 2022.

A. Radford and K. Narasimhan. Improving language understanding by generative pre-training. 2018.

C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67, 2020.

T. L. Scao, A. Fan, C. Akiki, E. Pavlick, S. Ilic, D. Hesslow, R. Castagné, A. S. Luccioni, F. Yvon, M. Gallé, J. Tow, A. M. Rush, S. Biderman, A. Webson, P. S. Ammanamanchi, T. Wang, B. Sagot, N. Muennighoff, A. V. del Moral, O. Ruwase, R. Bawden, S. Bekman, A. McMillan-Major, I. Beltagy, H. Nguyen, L. Saulnier, S. Tan, P. O. Suarez, V. Sanh, H. Laurençon, Y. Jernite, J. Launay, M. Mitchell, C. Raffel, A. Gokaslan, A. Simhi, A. Soroa, A. F. Aji, A. Alfassy, A. Rogers, A. K. Nitzav, C. Xu, C. Mou, C. Emezue, C. Klamm, C. Leong, D. van Strien, D. I. Adelani, and et al. BLOOM: A 176b-parameter open-access multilingual language model. *CoRR*, abs/2211.05100, 2022.

C. Schuhmann, R. Vencu, R. Beaumont, R. Kaczmarczyk, C. Mullis, A. Katta, T. Coombes, J. Jitsev, and A. Komatsuzaki. LAION-400M: open dataset of clip-filtered 400 million image-text pairs. *CoRR*, abs/2111.02114, 2021.

P. Sharma, N. Ding, S. Goodman, and R. Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *ACL (1)*, pages 2556–2565. Association for Computational Linguistics, 2018.

Y. Shen, K. Song, X. Tan, D. Li, W. Lu, and Y. Zhuang. Hugginggpt: Solving AI tasks with chatgpt and its friends in huggingface. *CoRR*, abs/2303.17580, 2023.

R. Taori, I. Gulrajani, T. Zhang, Y. Dubois, X. Li, C. Guestrin, P. Liang, and T. B. Hashimoto. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca, 2023.

H. Touvron, T. Lavril, G. Izacard, X. Martinet, M. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, and G. Lample. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971, 2023.

Vicuna. Vicuna: An open chatbot impressing gpt-4. https://github.com/lm-sys/FastChat, 2023.

Y. Wang, Y. Kordi, S. Mishra, A. Liu, N. A. Smith, D. Khashabi, and H. Hajishirzi. Self-instruct: Aligning language model with self generated instructions. *CoRR*, abs/2212.10560, 2022. doi: 10.48550/arXiv.2212.10560. URL https://doi.org/10.48550/arXiv.2212.10560.

C. Wu, S. Yin, W. Qi, X. Wang, Z. Tang, and N. Duan. Visual chatgpt: Talking, drawing and editing with visual foundation models. *CoRR*, abs/2303.04671, 2023.

C. Xu, D. Guo, N. Duan, and J. J. McAuley. Baize: An open-source chat model with parameter-efficient tuning on self-chat data. *CoRR*, abs/2304.01196, 2023a.

H. Xu, M. Yan, C. Li, B. Bi, S. Huang, W. Xiao, and F. Huang. E2E-VLP: end-to-end vision-language pre-training enhanced by visual learning. In *ACL/IJCNLP (1)*, pages 503–513. Association for Computational Linguistics, 2021.

H. Xu, Q. Ye, M. Yan, Y. Shi, J. Ye, Y. Xu, C. Li, B. Bi, Q. Qian, W. Wang, G. Xu, J. Zhang, S. Huang, F. Huang, and J. Zhou. mplug-2: A modularized multi-modal foundation model across text, image and video. *CoRR*, abs/2302.00402, 2023b.

Z. Yang, L. Li, J. Wang, K. Lin, E. Azarnasab, F. Ahmed, Z. Liu, C. Liu, M. Zeng, and L. Wang. MM-REACT: prompting chatgpt for multimodal reasoning and action. *CoRR*, abs/2303.11381, 2023.

Q. Ye, G. Xu, M. Yan, H. Xu, Q. Qian, J. Zhang, and F. Huang. Hitea: Hierarchical temporal-aware video-language pre-training. *CoRR*, abs/2212.14546, 2022. doi: 10.48550/arXiv.2212.14546. URL https://doi.org/10.48550/arXiv.2212.14546.

S. Zhang, S. Roller, N. Goyal, M. Artetxe, M. Chen, S. Chen, C. Dewan, M. T. Diab, X. Li, X. V. Lin, T. Mihaylov, M. Ott, S. Shleifer, K. Shuster, D. Simig, P. S. Koura, A. Sridhar, T. Wang, and L. Zettlemoyer. OPT: open pre-trained transformer language models. *CoRR*, abs/2205.01068, 2022.

D. Zhu, J. Chen, X. Shen, X. Li, and M. Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models, 2023a.

W. Zhu, J. Hessel, A. Awadalla, S. Y. Gadre, J. Dodge, A. Fang, Y. Yu, L. Schmidt, W. Y. Wang, and Y. Choi. Multimodal C4: an open, billion-scale corpus of images interleaved with text. *CoRR*, abs/2304.06939, 2023b.

A Training Hyperparameters

We report the detailed model training hyperparameters for visual knowledge learning in Table 4 and vision-language joint instruction tuning in Table 5.

Hyperparameters	
Training steps	50,000
Warmup steps	375
Max length	512
Batch size of image-caption pairs	4,096
Optimizer	AdamW
Learning rate	2e-4
Learning rate decay	Cosine
Adam ϵ	1e-6
Adam β	(0.9, 0.98)
Weight decay	0.01

Table 4: Training hyperparameters for multi-modal pre-training stage.

Hyperparameters	
Training steps	2,000
Warmup steps	50
Max length	1,024
Batch size of text instruction data	128
Batch size of multi-modal instruction data	128
Optimizer	AdamW
Learning rate	2e-5
Learning rate decay	Cosine
AdamW ϵ	1e-6
AdamW β	(0.9, 0.999)
Weight decay	0.0001

Table 5: Training hyperparameters for vision-language joint instruction tuning stage.

B Comparison with MM-REACT

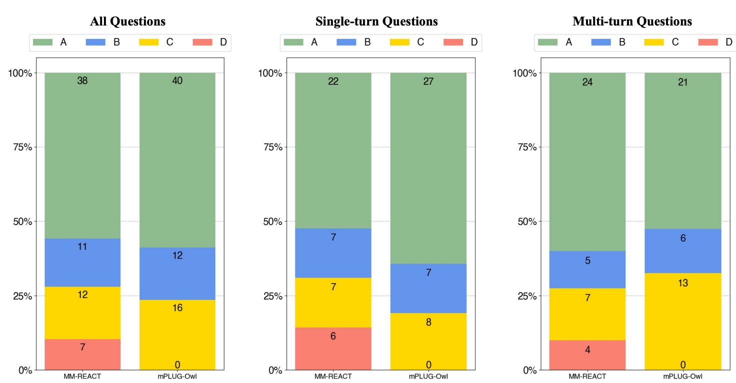
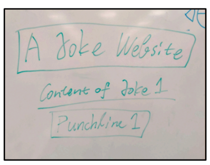
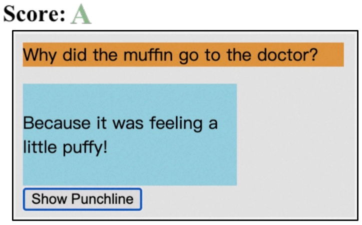


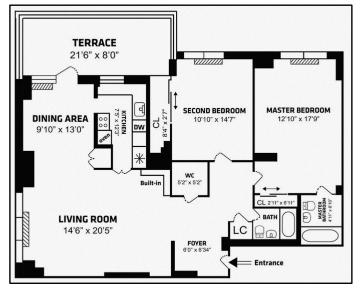
Figure 15: The comparison results which exclude the cases that were generated unsuccessfully by MM-REACT.



Write brief HTML/JS to turn this mock-up into a colorful website, where the jokes are replaced by a real joke

Score: A





How many bedrooms are there in this floor plan?

Score: A

There are two bedrooms in this floor plan.

What is the dimension of Master Bedroom?

Score: C

12' x 10'

Is there a bathroom in master bedroom?

Score: A

Yes, there is a bathroom in master bedroom.

What is the dimension of Second Bedroom?

Score: C

11' x 8'

Is the terrace bigger than the kitchen?

Score: A

Yes, the terrace is bigger than the kitchen.

What are the appliances in the kitchen in this floorplan?

Score: A

Stove, Refrigerator, Microwave, Dishwasher

What will you see when you enter through the entrance? and what is the dimension of it?

Score: B

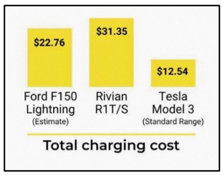
When you enter through the entrance, you will see the kitchen with a large refrigerator in it. The entrance doorway is 10 feet long, which gives an idea of the size and layout of this space. This is a floor plan for a new house, and the dimensions of the kitchen and the entrance doorway suggest a larger, more spacious living area.

Can you summarize this floorplan based on the aspects I was interested in, in a few sentences?

Score: B

A spacious floor plan with two main living areas, a large living room and a dining room, both 15 feet wide.
A master bedroom with a size of 12 feet by 10 feet and a bathroom.
A terrace bigger than the kitchen, offering additional living space and a great view.
A kitchen with basic appliances, such as a stove, refrigerator, microwave, and dishwasher.

Figure 16: OCR of simple scenes (mostly scenes with few numbers and no calculation a).



Excluding Ford F150 Lightning, what is the average total charging cost?

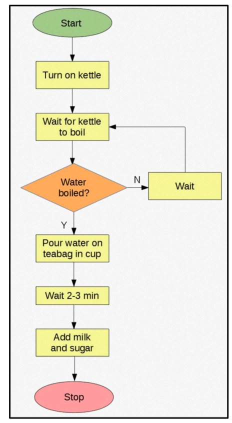
Score: C

The average total charging cost for the three electric cars, including the Ford F150 Lightning, can be calculated by multiplying the cost per kWh of electricity for each car, as provided in the image, by the number of miles driven. For example, the Ford F150 Lightning consumes 12 kWh per mile, the Chevy Bolt consumes 35 kWh per mile, and the Tesla Model S consumes 36 kWh per mile. The total charging cost would then be: ($1.02 per kWh for the Ford F150 Lightning) x (12 kWh/mile for the Ford F150 Lightning) x (200 miles for the Ford F150 Lightning) = $24.24. For the Chevy Bolt, it would be ($1.49 per kWh for the Chevy Bolt) x (35 kWh/mile for the Chevy Bolt) x (200 miles for the Chevy Bolt) = $1,138.40, and for the Tesla Model S, it would be ($0.94 per kWh for the Tesla Model S) x (36 kWh/mile for the Tesla Model S) x (200 miles for the Tesla Model S) = $316.48. These costs are estimates based on the information available in the image, and the actual charging costs may differ depending on factors like local electricity prices and driving habits.

Now how much cheaper is the average charging cost over the electric cars (including Ford F150 Lightning) than the average fueling cost over the gasoline cars (including Ford F150(Regular Cab))?Let's think step-by-step.

Score: C

Calculate the average total fueling cost of the gasoline cars:
Ford F150 Regular Cab: Average fueling cost is $1,240 per year
Ford F150 Regular Cab (EcoBoost): Average fueling cost is $1,324 per year
Ford F150 SuperCrew (4x2): Average fueling cost is $1,360 per year
Ford F150 SuperCrew (4x4): Average fueling cost is $1,396 per year
Ford F150 SuperCrew EcoBoost (4x2): Average fueling cost is $1,408 per year
Calculate the average total charging cost of the electric cars:
Ford F150 Lightning: Average charging cost is $5,400
Rivian R1T: Average charging cost is $2,490
Tesla Model 3: Average charging cost is $2,490
Chevy Bolt: Average charging cost is $2,490
Nissan Leaf: Average charging cost is $1,484
Calculate the difference between the two average total costs:
Gasoline cars: Average total fueling cost = $37,560 ($1,240+$1,324+$1,360+$1,396+$1,408)
Electric vehicles: Average total charging cost = $47,340 ($5,400+$2,490+$2,490+$2,490+$1,484)
The average total fueling cost for the gasoline cars is $37,560, and the average total charging cost for the electric vehicles is $47,340. Therefore, the electric vehicles have a total saving of $9,780 over 5 years, or about $1,958 per year.
This means that, on average, the electric vehicles have a savings of about $1,958 per year when compared to the gasoline-powered vehicles. This is due to the reduced fuel cost and maintenance associated with electric cars.



Can you read the text in this flowchart and explain it to me?

Score: B

The flowchart displays instructions for preparing a drink, specifically a mixture of orange juice and milk, which are placed in a blender. The process consists of four steps - slicing the oranges, putting them in the blender, adding milk, and then turning on the blender. The final step instructs to serve the blended drink in a glass. This simple and straightforward flowchart provides a clear guide for the person following the instructions to create the orange and milk drink easily and efficiently.

Figure 17: OCR of complex scenes (a).

Is there text in the image? If so, can you read it and explain it to me?

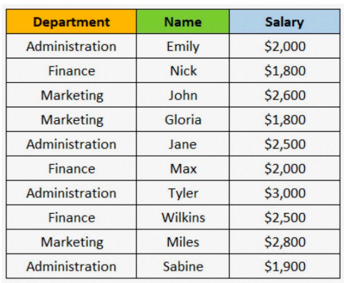
Score: C

Yes, there is text in the image. The text reads "Before... I don't feel good about my looks, test the text, am I pregnant? Now... I don't feel good about my looks, take the test, do I have the con?" This message is a joke about a person's insecurities about their appearance and the stress they feel about taking a pregnancy test.

What is funny about it?

Score: C

The joke is funny in several ways. First, it uses contrast to highlight the difference in the person's feelings about their appearance in the past and the present. The past self is unsure, while the present self seems more certain. This creates humor as the reader might relate to similar feelings of uncertainty about one's appearance.
Second, the message also makes fun of the stress people often feel when taking pregnancy tests. Many individuals might experience anxiety or worry when taking such tests, especially if they are not trying to get pregnant. The juxtaposition of these two elements, the person's appearance insecurities and the pregnancy-test anxiety, adds to the humor of the image and creates a relatable scenario for many individuals.



Department	Name	Salary
Administration	Emily	$2,000
Finance	Nick	$1,800
Marketing	John	$2,600
Marketing	Gloria	$1,800
Administration	Jane	$2,500
Finance	Max	$2,000
Administration	Tyler	$3,000
Finance	Wilkins	$2,500
Marketing	Miles	$2,800
Administration	Sabine	$1,900

How much is Wilkins earning?

Score: C

Wilkins is earning $110,000 per year.

Which department is the person who has the highest salary from?

Score: C

The person with the highest salary from is in the sales department.

This table is reporting monthly salary,can you calculate the yearly salary for each employee, and summarize it into a table and group them by department?

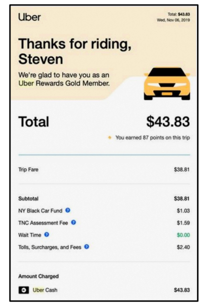
Score: B

Yes, I can provide a table summarizing the yearly salaries for each employee based on the monthly salary information provided. The table can be arranged in the following manner:\n\n| Name | Department | Monthly Salary | Yearly Salary |\n| ---- | ---- | ---------- | ----------- |\n| Wilkins | Sales | $110,000 | $1.32 million |\n| Smith | Sales | $90,000 | $1.08 million |\n| Jones | Sales | $68,000 | $836,000 |\n| Brown | Sales | $50,000 | $600,000 |\n| White | Engineering | $75,000 | $875,000 |\n| Black | Engineering | $85,000 | $1.03 million |\n| Grey | Engineering | $60,000 | $720,000 |\n| Green | Engineering | $45,000 | $540,000 |\n| Blue | Engineering | $60,000 | $720,000 |\n\nBy arranging the data in this table, we can see a clear breakdown of the yearly salaries for each employee and their respective departments. We can also see which departments have the highest and lowest average yearly salary.\n\nIn conclusion, the image of four employees sitting at a table displaying their monthly and yearly salaries, along with their department, provides a quick and informative overview of the salary distribution in the company.

Given the yearly bonus is 10%,12% and 15%for Administration,Finance,Marketing, respectively. Please provide me with the bonus for Emily, Max and Gloria.

Score: C

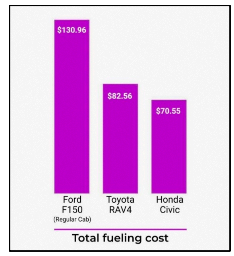
Based on the information provided, the bonuses for Emily, Max, and Gloria are 10%, 12%, and 15% respectively. This bonus is calculated as 0.10 x (100-500)/100 = 10%, 12% x (100-750)/100 = 12%, and 15% x (100-1000)/100 = 15% of their base salary.



What is the total amount?

Score: C

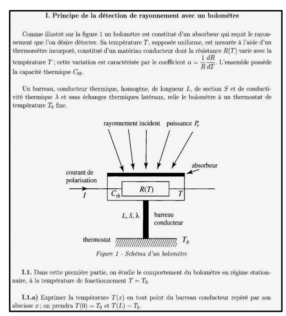
The total amount of money in the image is $33.83.



Excluding the regular cab, what is the average total fueling cost?

Score: C

The average total fueling cost for the vehicles in the image is $10,000. This is calculated by adding all the total fueling costs for each vehicle and then taking the average. The cost per gallon for each vehicle varies, with Ford F-150 costing $2.35 per gallon, Ford F-450 costing $2.44 per gallon, Ford Ranger costing $2.48 per gallon, and Ford C-Max Hybrid costing $2.58 per gallon. When these costs are added together to get the total, the average fueling cost for all four vehicles is $2.50 per gallon, which is $10,000 for 40,000 gallons.



I. Principe de la détection de rayonnement avec un bolomètre

Comme illustré sur la figure 1 un bolomètre est constitué d'un absorbeur qui reçoit le rayonnement que l'on désire détecter. Sa température T, supposée uniforme, est mesurée à l'aide d'un thermomètre incorporé, constitué d'un matériau conducteur dont la résistance $R(T)$ varie avec la température T ; cette variation est caractérisée par le coefficient $\alpha = \frac{1}{R}\frac{dR}{dT}$. L'ensemble possède la capacité thermique C_{th}.

Un barreau, conducteur thermique, homogène, de longueur L, de section S et de conductivité thermique λ et sans échanges thermiques latéraux, relie le bolomètre à un thermostat de température T_b fixe.

Figure 1 - Schéma d'un bolomètre

I.1. Dans cette première partie, on étudie le comportement du bolomètre en régime stationnaire, à la température de fonctionnement $T = T_0$.

I.1.a) Exprimer la température $T(x)$ en tout point du barreau conducteur repéré par son abscisse x ; on prendra $T(0) = T_0$ et $T(L) = T_b$.

Answer question I.1.a. Think step-by-step.

Score: C

To calculate the total heat loss of the heater, we first need to determine its thermal conductivity. This can be done by dividing the thermal conductivity of the material of the heater by its thickness.
The thermal conductivity of the metal of the heater is usually given in units of W/m-K. To calculate the thermal conductivity of the heater using this unit, we need to divide the value by the thickness in millimeters. The metal of the heater is typically made of steel, and its thickness is 5 mm. Therefore, its thermal conductivity is calculated to be 150 W/m-K.
Next, we need to find the heat flow from the heater to the environment, which is represented by the RTD. We know that the RTD has a resistance of R = 100 Ohms and a thermal resistance of Rt = 0.01 Ohm-C.
To calculate the total heat loss of the heater, we multiply the power input of the heater, which is given by P = 1 kW, by the RTD resistance and the thermal resistance. The result is 10.0 W of heat loss from the heater to the environment through Rt and the RTD resistance.
Finally, we can use the formula for heat loss through thermal conductivity to calculate the total heat loss of the heater:
Total heat loss of the heater = 0.02 W/m-K * 5 mm^2 * 1 kW = 10 W.
Therefore, the total heat loss of the heater is 10 W.

Figure 18: OCR of complex scenes (b).