Can Large Multimodal Models Uncover Deep Semantics Behind Images?

Anonymous ACL submission

Abstract

Understanding the deep semantics of images is essential in the era dominated by social media. However, current research works primarily on the superficial description of images, revealing a notable deficiency in the systematic investigation of the inherent deep semantics. In this work, we introduce DEEPEVAL, a comprehensive benchmark to assess Large Multimodal Models' (LMMs) capacities of visual deep semantics. DEEPEVAL includes human-annotated dataset and three progressive subtasks: fine-grained description selection, in-depth title matching, and deep semantics understanding. Utilizing DEEPEVAL, we evaluate 9 open-source LMMs and GPT-4V(ision). Our evaluation demonstrates a substantial gap between the deep semantic comprehension capabilities of existing LMMs and humans. For example, GPT-4V is 30% behind humans in understanding deep semantics, even though it achieves human-comparable performance in image description. Further analysis reveals that LMM performance on DEEPEVAL varies according to the specific facets of deep semantics explored, indicating the fundamental challenges remaining in developing LMMs.

1 Introduction

The Image is more than an idea. It is a vortex or cluster of fused ideas and is endowed with energy.

— Ezra Pound (1915)

Deep semantics of an image refer to the underlying meanings that extend beyond the superficial interpretation, probing into the essence of the image (Barthes, 1968). Although not every image inherently carries profound semantics, the concept of deep semantics is widespread across various fields (Barthes, 1999; Deman, 2010; Barthes, 2000; Somov, 2005, 2006). Understanding the deep semantics of images is a manifestation of high-level human intelligence, serving as an important means of exploration from perceptual intelligence to cognitive intelligence.

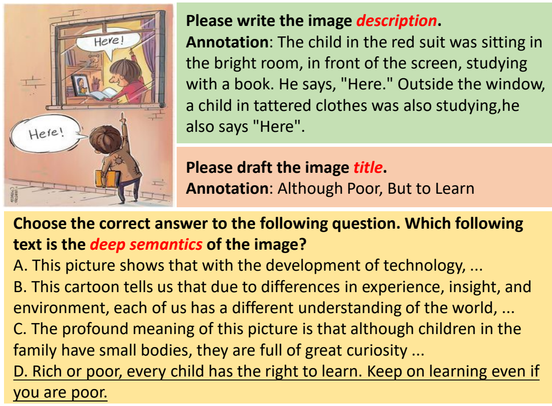


Figure 1: An example from the DEEPEVAL dataset includes annotated description, annotated title, and the corresponding multiple-choice question for deep semantics from the deep semantics understanding task.

However, previous efforts in visual understanding mainly focus on surface-level aspects of images, such as object attributes (Wang et al., 2022) and relationship reasoning (Hudson and Manning, 2019). Earlier attempts on deep semantic are limited in scope, focusing solely on sarcasm or humor, (Cai et al., 2019a; Chauhan et al., 2022; Boccignone et al., 2017; Patro et al., 2021), and lack in systematic investigation of the inherent deep semantic.

To address the mentioned limitations and fill the current research gap, we introduce DEEPEVAL, a benchmark for understanding the deep semantics of cartoons across various categories, accompanied by a meticulously annotated dataset. Additionally, we have devised three tasks: Fine-grained Description Selection, In-depth Title Matching, and Deep Semantics Understanding, to comprehensively evaluate models' capabilities in understanding deep semantics. Cartoons, often imbued with profound meanings by their creators, are an ideal subject for this study. The DEEPEVAL dataset comprises over

1,000 samples, each featuring a cartoon image and manually annotated components, including image description, title, and deep semantics. Moreover, we have developed multiple-choice questions for quantitative assessment, tailored for each task.

We conduct evaluations on various open-source LMMs as well as the proprietary GPT-4V(ision). Our findings reveal a significant gap between the capabilities of AI and humans in understanding deep meaning. Models with a larger number of parameters generally demonstrate a better understanding of deep semantics. Moreover, we discover that incorporating a description significantly helps these models in grasping the underlying semantics of an image. Furthermore, We also explore the performance of different models across various categories of images. By undertaking this task, we want to promote research in the development of models capable of comprehending deep semantics within visual content.

2 Related Work

Large Multimodal Models Large language models (LLMs) have exhibited strong abilities in various natural language understanding and generation tasks (Touvron et al., 2023a,b; Ray, 2023). Drawing on LLMs' scaling law, a series of Large Multimodal Models (LMMs) using LLMs as the backbone has emerged. These models (Tsimpoukelli et al., 2021; Li et al., 2023c; Liu et al., 2023b,a; Zhu et al., 2023; Wang et al., 2023; Ye et al., 2023) have aligned visual features with language models through additional layers or specialized modules. Several closed-source LMMs (Alayrac et al., 2022; Driess et al., 2023), especially GPT-4V (Yang et al., 2023b), show remarkable ability in managing complex multimodal inputs. These models have set new benchmarks in performance (Fu et al., 2023; Li et al., 2023a), increasingly becoming predominant in visual-language research. However, relevant studies suggest that LMMs still face limitations in comprehending deeper semantics (Liu et al., 2023c; Yang et al., 2023a).

Visual Deep Semantics Understanding Understanding the deep semantics of visual content represents a critical cognitive ability in humans. For AI, this ability showcases its depth of understanding images (Wang et al., 2021; Kruk et al., 2023). Present evaluations (Lin et al., 2014; Antol et al., 2015; Goyal et al., 2017; Gurari et al., 2018; Hudson and Manning, 2019; Wang et al., 2022; Xia et al., 2023) mainly concentrate on superficial aspect of understanding. Pioneering research in affective image classification (Yanulevskaya et al., 2008; Machajdik and Hanbury, 2010) has shown that LLMs are capable of attaining an understanding beyond mere surface content. Research in sarcasm (Das and Clark, 2018; Cai et al., 2019b; Lemmens et al., 2020; Abu Farha et al., 2022) and humor detection (Radev et al., 2016) only employs classification tasks. The further work (Desai et al., 2022) provides explanations for satirical content. The most relevant prior work Hessel et al. (2022) select humorous captions for images and provide explanations. However, it exclusively focuses on humor evaluation. In contrast, our work is pioneering in its comprehensive exploration of visual deep semantics across multiple categories, offering a more thorough assessment of the deep semantics within images. We provide a detailed comparison between our method and previous studies in Table 1, and the categories covered by our method are illustrated in Figure 2.

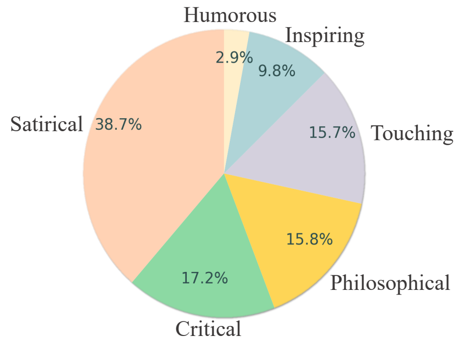


Figure 2: The distribution of six categories of DeepEval dataset.

3 Dataset and Task overview

The DeepEval dataset includes 1,001 samples, each with an image and three manually annotated components: a description, a title, and deep semantics, with the statistical information about the text displayed in Table 2. To enable quantitative evaluation, we additionally crafted multiple-choice questions to test the understanding of descriptions, titles, and deep semantics. Each segment is represented by 1,001 questions, where each question presents an image, a question text, and four potential answers. Only one answer is correct, while the others serve as distractors. Figure 1 illustrates examples of the manually annotated components and the multiple-choice questions.

Benchmark	Task	Semantics avg. length	Semantics size	# Category	Img Type
HCD (Radev et al., 2016)	Funniness Classification	-	-	1	
FSD (Das and Clark, 2018)	Sarcasm Classification	-	-	1	
MTSD (Cai et al., 2019b)	Sarcasm Classification	-	-	1	
RTSD (Lemmens et al., 2020)	Sarcasm Classification	-	-	1	
iSarcasmEval (Abu Farha et al., 2022)	Sarcasm Classification	-	-	1	
MORE (Desai et al., 2022)	Sarcasm Explanation	15	3510	1	
HUB(Hessel et al., 2022)	Matching+Ranking+Explanation	60	651	1	
DEEPEVAL(Ours)	Description+Title+**Deep Semantics**	37	1001	6	

Table 1: Features and statistical information of DEEPEVALand prior related datasets. "Semantics" refers to the explanatory texts in More and HUB, as well as annotated deep semantics texts in our dataset. "Img Type" includes black and white images and color images. The "-" refers to no semantics text in classification task.

Dataset Size		Description Length	
		tot.	avg.
1001		49,595	49.55
Deep Semantics Length		Title Length	
tot.	avg.	tot.	avg.
37,002	36.97	5,709	5.70

Table 2: Statistics of DEEPEVALdataset. The length of the text is calculated by counting the number of words contained in the text.

To explore the capabilities of LMMs in comprehending the deep semantics of image, we construct a comprehensive evaluation evaluation consisting of three main subtasks:

- *Fine-grained Description Selection Task:* Evaluating the ability of models to accurately identify the surface-level details of images.
- *In-depth Title Matching Task:* Assessing the capability of models to understand the overall signification of images.
- *Deep Semantics Understanding Task:* Evaluating the ability of models to understand the detailed deep semantic meanings of images.

Together, these subtasks offer a robust and multifaceted evaluation of LLMs, enabling a deeper understanding of their strengths and limitations in image understanding.

4 Dataset Construction

We collect DEEPEVAL dataset in a multi-step crowd-sourcing pipeline, including 1) image collection, 2) data annotation, 3) options generation. With selected high-quality comic images, we ask crowd-source workers to write a description, a title and deep semantics of each image.

4.1 Image collection

The image data in the DEEPEVAL dataset were obtained by web scraping from websites. A total of 1001 images were collected from Pinterest[1], Cartoon Movement[2], and Google search. The gathered images span a diverse array of genres, encompassing satirical representations of current events, philosophical narratives, humorous and entertaining content, among others. After collection, a manual screening process was conducted to remove duplicates and unclear images.

4.2 Data annotation

Deep semantics of images often requires extensive common sense knowledge and advanced reasoning abilities. To obtain high-quality image descriptions and connotation interpretations, we primarily utilize manual annotation to collect gold-standard answers with rigorous quality controls.

Annotator Recruitment and Instruction We posted a job description on online forums to invite over 50 applicants with at least a Bachelor's degree to participate in an online pre-annotation instruction and qualification test. Based on their preferences, we divided them into two groups: annotators and inspectors. After completing the pre-annotation instructions, we conducted a qualification test for quality control. In the end, we selected 26 annotators and 18 inspectors.

Cross-checking Annotation We divide the annotation process into 3 phases. In the first phase, annotators randomly selected comic images from the dataset for annotation of image description, title and deep semantics. The image description and deep semantics should be over 80 characters,

[1]https://www.pinterest.com/
[2]https://cartoonmovement.com/

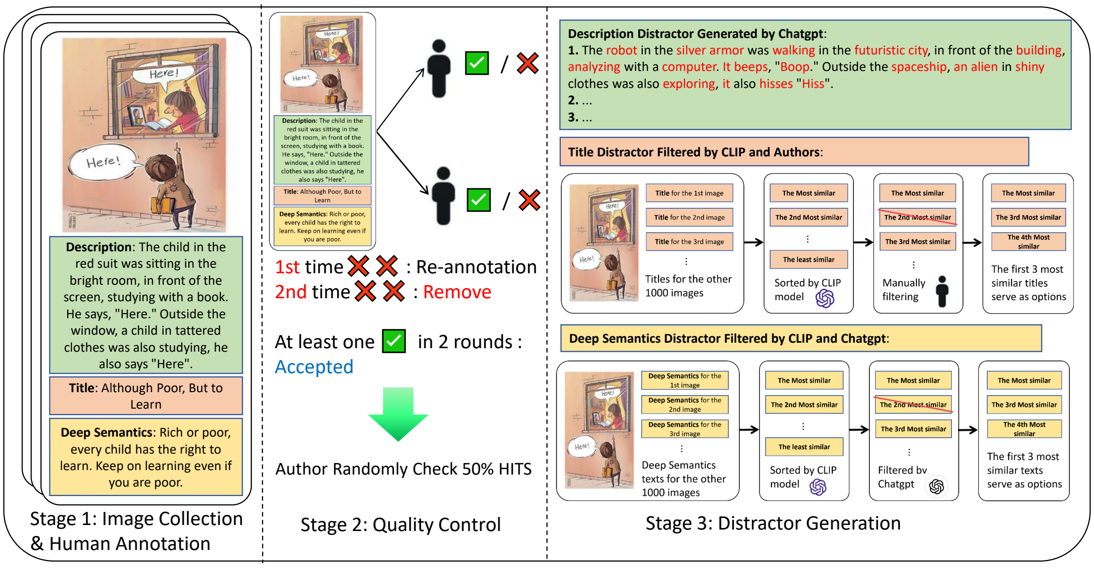


Figure 3: Schematic diagram of DEEPEVAL dataset construction process including three stages: Image Collection & Human Annotation, Quality Control and Distractor Generation.

while the proposed title should be over 3 characters, or else they cannot be submitted. Subsequent to this phase, each image has been transformed into a quadruple (image, description, title, deep semantics), marking the completion of the initial dataset construction.

In the second phase, inspectors will check the annotated images. When encountering low-quality annotations, they can drop them. Each image annotation will be checked by two inspectors. If both inspectors drop the annotation, we will drop it and put the image back into the dataset for annotation. If a comic image is rejected in two rounds of annotation, it means that the semantics conveyed by the image is not clear, and we will drop the image. At the culmination of this stage, we have elevated the quality of the dataset, essentially completing the foundational construction of the dataset.

In the third phase, to ensure the work of the inspectors, the author randomly selected 50% of the HITS from the second phase to ensure that the annotation meets our standards. In the final stage, we have acquired 1001 high-quality data entries, each represented as a quadruple (image, description, title, deep semantics).

4.3 Options Generation

After obtaining the image annotations, we use the annotated text as the correct option and construct three distractor options. Considering the high cost of constructing all distractor options using manual annotators, we have utilized the power of CLIP model and ChatGPT model in this section.

For the image description, we employ the ChatGPT model to generate sentences that retain sentence format but change the nouns, verbs, adverbs, or adjectives in the text to generate more intrusive options in the detailed description. The author manually checked all options in 100 images to ensure that the multiple-choice questions maintain a certain level of difficulty while having a unique and correct answer. Noting that detailed prompt and examples can be found in Appendix A.

For the deep semantics of the image, we use the CLIP model to calculate the similarity between the image and other underlying semantics texts. We aim to select texts with higher similarity scores as distractors to create more challenging options. However, due to the presence of images with similar themes in the dataset, which may share similar semantics and potentially cause confusion, we utilize the ChatGPT model to eliminate distractor texts that are too similar to the correct option. Subsequently, we select the top three terms with the highest similarity as distractor terms.

For image titles, we similarly utilize the CLIP model to determine the similarity between the image and other titles. However, as there may be numerous titles with distinct semantics that could potentially serve as the title for the same image, discerning whether a title causes confusion becomes more challenging. Therefore, in this part, the authors manually filter out confusing distractor texts and select texts with high similarity scores as distractor options, thus creating a challenging task that minimizes confusion.

4.4 Subtask Composition

We divide the task of understanding the deep semantics of cartoon into three progressive parts: fine-grained image description selection, in-depth title matching, and deep semantics understanding. Among them, the fine-grained description selection task requires multi-modal models to identify the surface-level details of the images. The in-depth title matching task requires models to comprehend the overall significance of the images and grasp their basic intentions. The deep semantics understanding task takes it a step further by demanding multi-modal models to acquire a comprehensive and detailed understanding of the thoughts, connotations, and information conveyed in the images. It can be observed that these three tasks gradually augment the comprehension of images, each task building upon the previous one to deepen the level of understanding. In these three tasks, each question consists of an image and a multiple-choice question with four options. The model is then required to choose the option it believes is the best description, title, or connotation from the four options.

4.5 Dataset quality

To ensure the quality of the dataset, the authors have randomly checked 50% of the data for descriptions, titles, deep semantics annotations, and the multiple-choice questions of the three tasks. this ensures that the content of descriptions, titles, and deep semantics annotations meet the standards and maintained high quality. For the multiple-choice questions, this confirms that they were challenging and had standard answers. Furthermore, we employed annotators to evaluate the triplets of each image (description, title, semantics) and provide a score between 1 and 5. A score of 1 indicates complete inconsistency, a score of 5 indicates complete consistency, and each image is evaluated by three different annotators. Finally, our dataset obtained an average score of 4.74, indicating that our dataset is of high quality.

4.6 License and Copyright

In this dataset, we used original web links of comic images without infringing on their copyright. For images sourced from MathPile governed by licenses stricter than CC BY-NC-SA 4.0, MathPile adheres to the more restrictive licensing terms. Otherwise, it operates under the CC BY-NC-SA 4.0 license. This work is licensed under a CC BY-NC license. Our annotators participated in the annotation process voluntarily and received fair compensation.

5 Experiments

5.1 Baselines

In consideration of the strong performance exhibited by Large Multimodal Models(LMMs) in addressing image comprehension challenges, we introduce the following seven LMMs: LLaVA-1.5 (Liu et al., 2023a), MiniGPT-4 (Zhu et al., 2023), mPLUG-Owl2 (Ye et al., 2023), CogVLM (Wang et al., 2023), Qwen-VL (Bai et al., 2023b), InstructBlip2 (Dai et al., 2023), Fuyu (Bavishi et al., 2023), GPT-4V (Yang et al., 2023b). A detailed introduction to these models can be found in the Appendix F.

5.2 Experiment Details

In evaluating performance for our tasks, accuracy serves as the primary metric. A model's answer is deemed correct when it aligns with the established standard answer. Accuracy is quantified by the ratio of the number of correct responses N_r to the total number of question N, expressed as N_r/N. Our task prompts commence with a topic specification, encompassing description, title, or deep semantics, succeeded by multiple-choice options A, B, C, and D. To minimize deviations in results caused by variations in the text descriptions within the prompt, we have developed three distinct prompt formats, which are elaborately described in Appendix refsec:prom-deta. The parameters for each model used in the experiment, including possible settings for temperature and top-k, are comprehensively detailed in Appendix refsec:mod-para-de. Furthermore, to assess human capabilities in these tasks, we randomly select 100 questions from the dataset for each task and have annotators answer them. This allows us to benchmark the performance of human participants against our models, offering a thorough comparison of both human and machine proficiency in these specific tasks.

5.3 Main Results

Fine-grained Description Selection Task The results of various multimodal large models in fine-grained description selection task are shown in

Model	Backbone	# Params	Description	Title	DeepSemantics
CogVLM (Wang et al., 2023)	Vicuna-v1.5	17B	$72.83_{\pm 6.81}$	$45.05_{\pm 5.89}$	$32.20_{\pm 1.00}$
InstructBlip-13B (Dai et al., 2023)	Vicuna-v1.5	14B	$59.44_{\pm 6.12}$	$36.66_{\pm 3.55}$	$15.75_{\pm 2.04}$
LLaVA-1.5-13B (Liu et al., 2023a)	Vicuna-v1.5	13B	$53.91_{\pm 10.92}$	$35.13_{\pm 5.16}$	$25.71_{\pm 0.16}$
Qwen-VL-Chat (Bai et al., 2023b)	Qwen	10B	$78.82_{\pm 4.68}$	$47.68_{\pm 1.79}$	$28.30_{\pm 0.40}$
mPlug-Owl2 (Ye et al., 2023)	LLaMA2	8B	$75.26_{\pm 3.66}$	$47.75_{\pm 0.85}$	$31.37_{\pm 2.55}$
MiniGPT-4 (Zhu et al., 2023)	LLaMA2	8B	$41.79_{\pm 5.74}$	$33.00_{\pm 4.30}$	$26.34_{\pm 2.24}$
InstructBlip-7B (Dai et al., 2023)	Vicuna-v1.5	8B	$49.88_{\pm 6.18}$	$32.23_{\pm 4.87}$	$15.72_{\pm 1.26}$
Fuyu (Bavishi et al., 2023)	-	8B	$27.04_{\pm 5.14}$	$26.60_{\pm 1.00}$	$20.21_{\pm 3.53}$
LLaVA-1.5-7B (Liu et al., 2023a)	Vicuna-v1.5	7B	$48.62_{\pm 13.61}$	$32.00_{\pm 6.48}$	$24.94_{\pm 2.05}$
GPT-4V (Yang et al., 2023b)	-	-	$\mathbf{96.53}_{\pm 1.06}$	$\mathbf{55.01}_{\pm 0.96}$	$\mathbf{63.14}_{\pm 3.00}$
Human	-	-	100.00	94.00	93.00

Table 3: The benchmark includes the average accuracy (in percentages (%)) and variance on three prompts for the DEEPEVALmethod. Description, Title and DeepSemantics represent Fine-grained Description Selection Task, In-depth Title Matching Task, and Deep Semantics Undertanding Task respectively.

Table 3. It can be observed that Qwen-VL-Chat, among the open-source models, exhibit the highest recognition capability for fine-grained surface description, with an accuracy of 78.82%. On the other hand, Fuyu demonstrates the weakest recognition ability for fine-grained surface-level information, with an accuracy of only 27.04%. The latest GPT-4V exhibits outstanding performance with an impressive accuracy of 96.53%. Nevertheless, these models still do not match the capabilities of humans, whose accuracy remains at a perfect 100%.

In-depth Title Matching Task The performance of the model in the in-depth title matching task is also presented in Table 3. Among the open-source models, mPlug-Owl2 performs the best with an accuracy of 47.75%, while Fuyu shows the weakest performance with an accuracy of only 26.60%. The closed-source model GPT-4V outperforms them all, achieving an accuracy of 55.01%. A notable observation across all models is that their performance in this task significantly trails behind their performance in the preceding fine-grained description selection task. This indicates that processing deep semantics is more challenging, despite the in-depth title matching task primarily addressing the broad essence rather than intricate details of deep semantics. Additionally, it's evident that these models substantially fall short of human-level performance, which is marked at an impressive 94%.

Deep Semantics Understanding Task Among open-source models, CogVLM showcases the highest performance with an accuracy of 32.20%, while LLava-1.5-7B scores the lowest, achieving only 15.72% accuracy, shown in Table 3. Unsurprisingly, GPT-4V achieves better results with an accuracy of 63.14%. However, GPT-4V exhibits the largest variance among all models, excluding Fuyu, in deep semantics understanding, indicating instability despite its overall superior performance. Furthermore, when comparing GPT-4V's results across all tasks, there is notably higher variance in the deep semantics aspect, suggesting weaker performance compared to other tasks. Additionally, we note that the capabilities of these models are significantly weaker than human performance, which stands at 93%.

It can be observed that the accuracy of all evaluated models in deep semantics understanding is significantly lower than their performance in image description, and nearly all of them achieve lower accuracy in deep semantics understanding compared to the in-depth title matching task. This underscores that comprehending the deep semantics of images presents a significant challenge for these models, and focusing on the finer details of deep semantics adds further complexity, aligning with our expectations. Interestingly, only GPT-4V demonstrates higher accuracy in the deep semantics task compared to the in-depth title matching task. This could suggest that GPT-4V's stronger understanding of longer texts, coupled with the detailed information conveyed in deep semantics texts, aids the model in making more accurate judgments in deep semantics understanding.

6 Analysis

6.1 How do models perform across various categories in image understanding?

By analyzing the model's understanding capabilities in different categories, we can pinpoint the

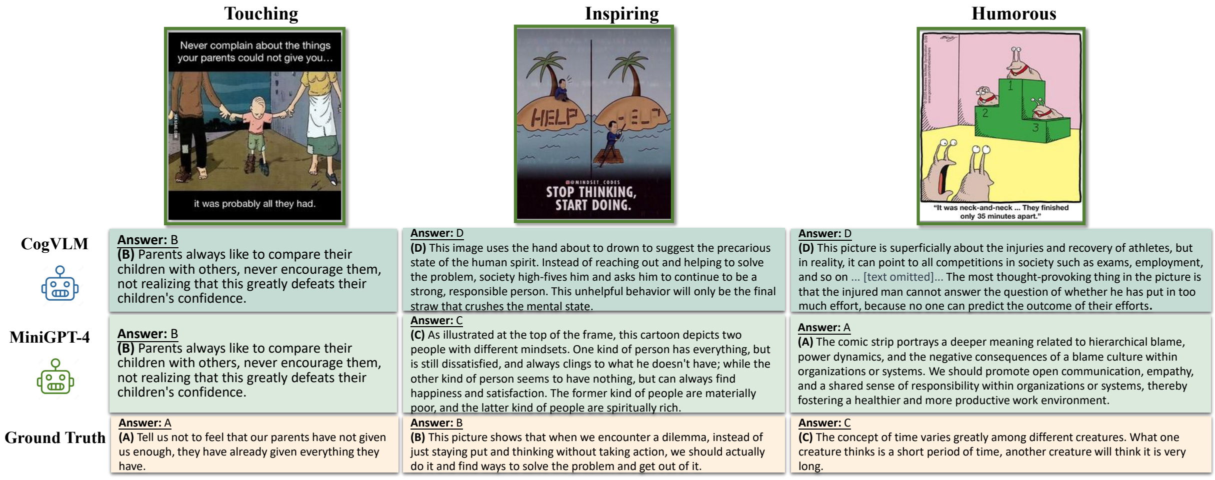


Figure 4: Random samples of answers chosen by CogVLM and MiniGPT-4, along with the standard answers, covering three categories: *Touching*, *Inspiring*, and *Humorous*, with one sample from each category.

specific categories of relative strength or weakness. The performance of different models across categories is illustrated in Figure 5, with three radar charts showcasing the model's ability in interpreting image descriptions, titles, and deep semantics across different categories. The deep semantics graph reveals that different models exhibit their strengths in different categories. For instance, the mPlug-Owl2 and CogVLM models stand out in the *Humorous* and *Inspiring* categories, respectively. Furthermore, despite extensive prior research, *Satirical* category continues to challenge all models, with accuracy rates remaining below 30%. This underscores the *Satirical* category as a critical area for further research in understanding deep semantics within images.

The Description Selection task's radar charts, resembling regular hexagons, indicate a more uniform comprehension of image descriptions across categories by the models. When evaluating titles, models show remarkable competency in both *Humorous* and *Inspiring* categories compared to others. However, regarding deep semantics, *Inspiring* consistently emerges as the top-performing category for four models, whereas a majority struggle with *Humorous*. This discrepancy may stem from the fact that *Inspiring* content often be succinctly summarized in few sentences, in contrast to *Humorous* content, which typically involves more intricate interpretations that are heavily reliant on cultural context, timing, and the subtleties of language and expression. To provide a more intuitive display, Figures 4 showcase samples from typical categories in the deep semantics understanding task for CogVLM, MiniGPT-4, and the standard answers, while additional samples for other categories are available in Figures 7 in the Appendix E .

6.2 Can image descriptions aid models' understanding of deep semantics?

It is commonly believed that models need to first identify the content of image descriptions before further comprehending the deep semantics. Therefore, we were curious to explore whether inspiring the model by incorporating its surface image descriptions during the evaluation process would aid in the model's understanding of deep semantics. This process is divided into two steps: 1) having the model to generate detailed descriptions of the images; 2) incorporating the detailed descriptions into the prompt of the deep semantics understanding task. Additionally, to more effectively demonstrate the impact of integrating image descriptions on the understanding of deep semantics, we also directly include annotated image description texts in the prompt. In this case, the first step is omitted, and the detailed descriptions included in the second step are the annotated detailed descriptions.

The results presented in Table 4 reveal that seven out of the evaluated models demonstrate an enhancement in their understanding of deep semantics upon the integration of model-generated image descriptions, with an average increase of 1.8 percentage points. Eight out of the evaluated models show enhanced deep semantics understanding capabilities when annotated image descriptions are added, with an average increase of approximately 4.1 percentage points. Thus, it appears that inspir-

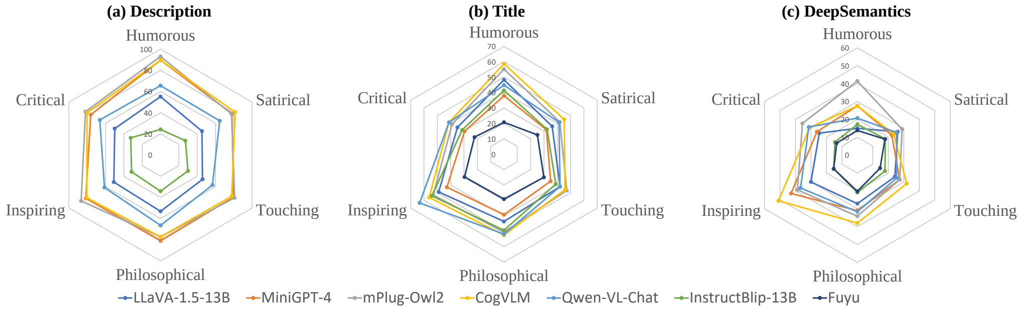


Figure 5: The radar charts represent the performance of several typical models in understanding images across different categories in our three tasks.

Model	**DS**	**DS (GeneDesc)**	**DS (AnnoDesc)**
CogVLM	31.17	32.57	37.96
InstructBlip-13B	17.77	19.78	23.48
LLaVA-1.5-13B	25.88	26.87	30.07
Qwen-VL-Chat	28.37	28.17	34.57
mPlug-Owl2	31.97	35.46	41.16
MiniGPT-4	27.27	27.77	34.07
InstructBlip-7B	14.29	19.38	19.38
Fuyu	16.98	16.78	23.98
LLaVA-1.5-7B	27.27	30.83	30.07

Table 4: The model's capability to comprehend the deep semantics of images while incorporating various image descriptions. "DS" stands for "Deep Semantics", "GeneDesc" represents model-generated image descriptions. "AnnoDesc" signifies annotated image descriptions.

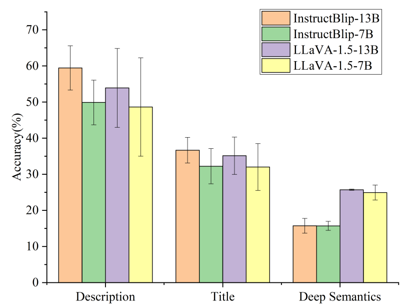


Figure 6: Comparison of the average accuracy and variance results between InstructBlip-13B vs InstructBlip-7B and LLaVA-1.5-13B vs LLaVA-1.5-7B.

ing the model by incorporating its descriptions of the surface content of images does contribute to enhancing the model's deep semantics understanding capabilities.

6.3 How does model parameter size affect deep semantics understanding?

Due to the Scaling Law, the number of parameters generally has a positive impact on the model's performance. In this context, we also discuss the relationship between model parameters size and deep understanding. We examine two pairs of models, InstructBlip-13B vs. InstructBlip-7B and LLaVA-1.5-13B vs. LLaVA-1.5-7B, where each pair has consistent architecture and training processes, differing only in parameter size. Figure 6 provide a visual representation of the mean and variance of accuracy across three tasks for these four models. It is observable that the 7B models have lower accuracy across all three tasks compared to the 13B models, indicating superior performance of the 13B models. Furthermore, the overall variance of the 7B models is higher than that of the 13B models. This indicates that, generally speaking, the 13B models are also more stable than the 7B models. From this, it is evident that an increase in the number of parameters indeed has a positive impact on the models' deep semantics understanding capabilities.

7 Conclusion

We propose DEEPEVAL, a benchmark for visual deep semantics of LMMs. DEEPEVALconsists of well-annotated dataset and three parts: fine-grained description selection task, in-deep title matching task, and deep semantic understanding task. Evaluations are conducted on the leading LLMs, revealing a significant gap between AI and human capabilities in understanding deep semantics. Further analysis indicates that integrating image descriptions during the inference process notably enhances LMMs' ability to perceive deep semantics. Existing models still have a long way to go in terms of visual deep semantics understanding compared to humans. We hope that the proposed dataset and tasks can pave the way for AI to achieve a deeper understanding of the profound semantics conveyed by images.

Limitations

The deep semantics of cartoon images are varied, and due to our limited collection of images, it's not feasible to encompass all potential deep semantic content. In this work, we have only exemplified some common categories, but the categories of images in the real world far exceed these six. On this note, adding more images and annotations would help improve this issue. Furthermore, our images currently only include cartoons. Future work can expand to incorporate more types of images, such as photographs, advertising images, and artworks. Lastly, in the annotation process, we aim to reach a consensus among annotators on the deep semantics of images and only retain images with agreed-upon deep semantics. Therefore, images with deep semantics but significant controversy will not be included.

References

Ibrahim Abu Farha, Silviu Vlad Oprea, Steven Wilson, and Walid Magdy. 2022. SemEval-2022 task 6: iSarcasmEval, intended sarcasm detection in English and Arabic. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 802–814, Seattle, United States. Association for Computational Linguistics.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023a. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023b. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*.

Roland Barthes. 1968. *Elements of semiology*. Macmillan.

Roland Barthes. 1999. Rhetoric of the image. *Visual culture: The reader*, pages 33–40.

Roland Barthes. 2000. The photographic message. *Theorizing communication: readings across traditions*, pages 191–199.

Rohan Bavishi, Erich Elsen, Curtis Hawthorne, Maxwell Nye, Augustus Odena, Arushi Somani, and Sağnak Taşırlar. 2023. Introducing our multimodal models.

Giuseppe Boccignone, Donatello Conte, Vittorio Cuculo, and Raffaella Lanzarotti. 2017. Amhuse: a multimodal dataset for humour sensing. In *Proceedings of the 19th ACM international conference on multimodal interaction*, pages 438–445.

Yitao Cai, Huiyu Cai, and Xiaojun Wan. 2019a. Multimodal sarcasm detection in twitter with hierarchical fusion model. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 2506–2515.

Yitao Cai, Huiyu Cai, and Xiaojun Wan. 2019b. Multimodal sarcasm detection in Twitter with hierarchical fusion model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2506–2515, Florence, Italy. Association for Computational Linguistics.

Dushyant Singh Chauhan, Gopendra Vikram Singh, Aseem Arora, Asif Ekbal, and Pushpak Bhattacharyya. 2022. An emoji-aware multitask framework for multimodal sarcasm detection. *Knowledge-Based Systems*, 257:109924.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Albert Li, Pascale Fung, and Steven C. H. Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. *ArXiv*, abs/2305.06500.

Dipto Das and Anthony J Clark. 2018. Sarcasm detection on flickr using a cnn. In *Proceedings of the 2018 international conference on computing and big data*, pages 56–61.

Jonathon Deman. 2010. The comics other: Charting the correspondence between comics and difference.

Poorav Desai, Tanmoy Chakraborty, and Md Shad Akhtar. 2022. Nice perfume. how long did you marinate in it? multimodal sarcasm explanation. In *Proceedings of the AAAI Conference on Artificial Intelligence (Volume 1: Long Papers)*, pages 10563—-10571. The Symposium on Educational Advances in Artificial Intelligence.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale.

Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. 2023. Palm-e: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*.

Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. 2022. Eva: Exploring the limits of masked visual representation learning at scale.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. 2023. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913.

Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. 2018. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3608–3617.

Jack Hessel, Ana Marasović, Jena D. Hwang, Lillian Lee, Jeff Da, Rowan Zellers, Robert Mankoff, and Yejin Choi. 2022. Do androids laugh at electric sheep? humor "understanding" benchmarks from the new yorker caption contest. *ArXiv*, abs/2209.06293.

Drew A Hudson and Christopher D Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709.

Julia Kruk, Caleb Ziems, and Diyi Yang. 2023. Impressions: Visual semiotics and aesthetic impact understanding. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12273–12291, Singapore. Association for Computational Linguistics.

Jens Lemmens, Ben Burtenshaw, Ehsan Lotfi, Ilia Markov, and Walter Daelemans. 2020. Sarcasm detection using an ensemble approach. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 264–269, Online. Association for Computational Linguistics.

Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. 2023a. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023b. Blip-2: Bootstrapping language-image pretraining with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. 2023c. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning*.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023a. Improved baselines with visual instruction tuning.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual instruction tuning.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. 2023c. Mmbench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*.

Jana Machajdik and Allan Hanbury. 2010. Affective image classification using features inspired by psychology and art theory. *Proceedings of the 18th ACM international conference on Multimedia*.

Badri N Patro, Mayank Lunayach, Deepankar Srivastava, Hunar Singh, Vinay P Namboodiri, et al. 2021. Multimodal humor dataset: Predicting laughter tracks for sitcoms. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 576–585.

Dragomir Radev, Amanda Stent, Joel Tetreault, Aasish Pappu, Aikaterini Iliakopoulou, Agustin Chanfreau, Paloma de Juan, Jordi Vallmitjana, Alejandro Jaimes, Rahul Jha, and Robert Mankoff. 2016. Humor in collective discourse: Unsupervised funniness detection in the new yorker cartoon caption contest. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 475–479, Portorož, Slovenia. European Language Resources Association (ELRA).

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision.

Partha Pratim Ray. 2023. Chatgpt: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet of Things and Cyber-Physical Systems*.

Noam Shazeer. 2020. Glu variants improve transformer.

Georgij Yu Somov. 2005. Semiotic systems of works of visual art: Signs, connotations, signals. *Semiotica*, 2005(157):1–34.

Georgij Yu Somov. 2006. Connotations in semiotic systems of visual art (through the example of works by ma vrubel).

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023b. Llama 2: Open foundation and fine-tuned chat models.

Maria Tsimpoukelli, Jacob L Menick, Serkan Cabi, SM Eslami, Oriol Vinyals, and Felix Hill. 2021. Multimodal few-shot learning with frozen language models. *Advances in Neural Information Processing Systems*, 34:200–212.

Lijuan Wang, Wenya Guo, Xingxu Yao, Yuxiang Zhang, and Jufeng Yang. 2021. Multimodal event-aware network for sentiment analysis in tourism. *IEEE MultiMedia*, 28(2):49–58.

Ruonan Wang, Yuxi Qian, Fangxiang Feng, Xiaojie Wang, and Huixing Jiang. 2022. Co-vqa: Answering by interactive sub question sequence. *arXiv preprint arXiv:2204.00879*.

Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, Jiazheng Xu, Bin Xu, Juanzi Li, Yuxiao Dong, Ming Ding, and Jie Tang. 2023. Cogvlm: Visual expert for pretrained language models.

Heming Xia, Qingxiu Dong, Lei Li, Jingjing Xu, Tianyu Liu, Ziwei Qin, and Zhifang Sui. 2023. ImageNetVC: Zero- and few-shot visual commonsense evaluation on 1000 ImageNet categories. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2009–2026, Singapore. Association for Computational Linguistics.

Xiaocui Yang, Wenfang Wu, Shi Feng, Ming Wang, Daling Wang, Yang Li, Qi Sun, Yifei Zhang, Xiaoming Fu, and Soujanya Poria. 2023a. Mmbigbench: Evaluating multimodal models on multimodal content comprehension tasks. *arXiv preprint arXiv:2310.09036*.

Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. 2023b. The dawn of lmms: Preliminary explorations with gpt-4v (ision). *arXiv preprint arXiv:2309.17421*, 9(1).

V. Yanulevskaya, J.C. van Gemert, K. Roth, A.K. Herbold, N. Sebe, and J.M. Geusebroek. 2008. Emotional valence categorization using holistic image features. In *2008 15th IEEE International Conference on Image Processing*, pages 101–104.

Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Anwen Hu, Haowei Liu, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. 2023. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.

A Examples of Generating Distractor Generation For Description

Considering that the generation of interference terms in the description only requires replacing nouns, adjectives, verbs, etc. in the sentence, we use Chatgpt to complete this task. The following is the prompt we use:*Give me three different paragraphs that take only some of the verbs, nouns, adjectives, and adverbs in a given paragraph and modify words with irrelevant meanings.*
Input: [Example Input 1]
Output: [Example Output 1]

Input: [Example Input 2]
Output: [Example Output 2]
Input: [Example Input 3]
Output: [Example Output 3]
Input: [Input]
Output:

To ensure that ChatGPT understands our requirements correctly, we use a 3-shot prompt. These three examples were manually written by the author. The following is a modification example. It should be noted that the output of each example in the prompt has 3 modified paragraphs of text. For convenience, only one modified paragraph of text is shown here

Source Text: *In the picture,there are three queues,the first one named Critic has many people,stand in an endless line;the second one named Talker also has many people,but not that much as Critic;the third queue named Doer,with no one in line.*

Revised Text: *In the picture, there are three cats, the first one named Critic has many toys, play in an endless loop; the second one named Talker also has many toys, but not that much as Critic; the third cat named Doer, with no toys to play with.*

B Prompt Details

To eliminate the influence of prompt expression on model performance, we used the following three types of prompts for testing:

- *Choose the correct answer to the following question. Which following text is the [description/best title/deep meaning] of the image?*
  Options: (A) [...] (B) [...] (C) [...] (D) [...]
  Answer:

- *Select the appropriate [description/title/deep meaning] for the image from the options given. Which of these is the most suitable [description/title/deep meaning] for the image?*
  Choices: A) [...] B) [...] C) [...] D) [...]
  Correct Answer:

- *Identify the most suitable [description/title/deep meaning] for the image from the given options. Which of the following should be chosen as the [description/title/deep meaning]?*
  Choices are: A. [...], B. [...], C. [...], and D. [...].
  The correct answer is:

C Model Hyper-parameter Details

We use the default hyper-parameter values of the models. In the LLaVa-1.5-7B and LLaVa-1.5-13B, the temperature is set to 0.2. For MiniGPT-4, the temperature is set to 1.0, and num_beams is also set to 1.0. The temperature for mPlug-Owl-2 is set to 0.7. For CogVLM, the temperature is set to 0.4, top_p is set to 0.8, and top_k is set to 1.0.

D Categories Definition

Table 5 give the names and detailed definitions of the categories in DEEPEVAL.

E Categories Samples

Figure 7 give the samples of answers chosen by CogVLM and MiniGPT-4, in three *Satirical*, *Critical*, and *Philosophical* category.

F Large Multimodal Models

- **LLaVA-1.5** (Liu et al., 2023a) is an end-to-end LMM extended from Vicuna(Chiang et al., 2023), augmented with vision encoder.

- **MiniGPT-4** (Zhu et al., 2023) is an extension of Vicuna, incorporating ViT (Dosovitskiy et al., 2021) and Q-former (Li et al., 2023b) as the vision encoder, while also featuring a single linear projection layer sandwiched between them.

- **mPLUG-Owl2** (Ye et al., 2023) is an extension of LLaMA-2-7B(Touvron et al., 2023b), using ViT-L/14(Radford et al., 2021) as the vision encoder, and introducing a visual abstractor between them.

- **CogVLM** (Wang et al., 2023) is also an extension of Vicuna, incorporating ViT(Dosovitskiy et al., 2021) as the vision encoder, a two-layer MLP(Shazeer, 2020) as adapter, and introducing Visual expert module.

- **Qwen-VL** (Bai et al., 2023b) is an extension of Qwen-7B(Bai et al., 2023a), incorporating ViT(Dosovitskiy et al., 2021) as the vision encoder, and introducing a vision-language adapter that compresses the image features.

- **InstructBlip2** (Dai et al., 2023) employs ViT-g/14 (Fang et al., 2022) as image encoder, and four different LLMs as language decoders. In

Table 5: The names and specific definitions of the categories in DEEPEVAL.

Category	Definition
Humorous	The image elicits amusement, laughter, or a sense of light-heartedness. It may contain elements that are funny, witty, or clever.
Critical	The image offers a critical perspective or analysis of a specific subject, aiming to examine and evaluate its merits, shortcomings, or implications.
Touching	The image evokes strong emotions such as joy, sadness, empathy, or nostalgia. It may depict a heartwarming scene, a tender moment, or a poignant event.
Philosophical	The image stimulates intellectual or philosophical contemplation. It raises questions, challenges assumptions, or encourages viewers to reflect on deeper meanings or concepts.
Inspiring	The image motivates or uplifts viewers, conveying a positive message, encouraging resilience, or instilling hope. It may depict acts of kindness, achievement, or triumph over adversity.
Satirical	The image conveys a message or commentary on a particular subject, often by using irony, sarcasm, or wit to highlight flaws or satirize societal norms, institutions, or individuals.

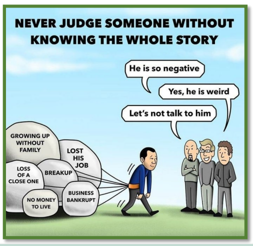
our following tests, we utilize vicuna-13B and vicuna-7B (Chiang et al., 2023) versions.

- **Fuyu** (Bavishi et al., 2023) employs a decoder-only architecture, devoid of a dedicated image encoder for image processing. This design choice enables the model to support arbitrary image resolutions.

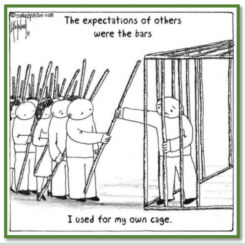
- **GPT-4V** (Yang et al., 2023b) is OpenAI's cutting-edge language model redefining natural language processing with advanced contextual understanding and versatile linguistic abilities.

Satirical

Critical



Philosophical



	Satirical	Critical	Philosophical
CogVLM	**Answer:** B **(B)** The cartoon expresses people's concern and anxiety about age. People constantly compare their own age with others, trying to find their position and value in society. However, such comparisons only make them feel more anxious and uneasy. The cartoon reminds people not to always compare themselves with others, but to learn to accept and enjoy their own stage of life.	**Answer:** D **(D)** As illustrated at the top of the frame, this cartoon depicts two people with different mindsets. One kind of person has everything, but is still dissatisfied, and always clings to what he doesn't have; while the other kind of person seems to have nothing, but can always find happiness and satisfaction. The former kind of people are materially poor, and the latter kind of people are spiritually rich.	**Answer:** B **(B)** The comic strip conveys a deeper meaning related to the divergent outcomes resulting from different choices and behaviors. The caricature also serves as a reminder that our actions have consequences and that making positive choices can lead to personal growth, fulfillment, and a vibrant life, while negative habits can lead to stagnation and a decline in overall well-being.
MiniGPT-4	**Answer:** B **(B)** The cartoon expresses people's concern and anxiety about age. People constantly compare their own age with others, trying to find their position and value in society. However, such comparisons only make them feel more anxious and uneasy. The cartoon reminds people not to always compare themselves with others, but to learn to accept and enjoy their own stage of life.	**Answer:** B **(B)** This cartoon exposes human nature. Even family members, no matter how good their previous relationships were, will turn their backs when it comes to financial interests.	**Answer:** B **(B)** The comic strip conveys a deeper meaning related to the divergent outcomes resulting from different choices and behaviors. The caricature also serves as a reminder that our actions have consequences and that making positive choices can lead to personal growth, fulfillment, and a vibrant life, while negative habits can lead to stagnation and a decline in overall well-being.
Ground Truth	**Answer:** C **(C)** This picture shows that some people are not on the same page, in order to maintain their own image choice and status, doing the crowning thing, while behind the scenes, but heartily despise the poor lower than their own status.	**Answer:** A **(A)** Without knowing a person's experience, we should not arbitrarily judge him or her, because this judgment is likely to be unfair	**Answer:** A **(A)** We live under the expectations of others, devoid of individual thoughts, which prevents us from becoming complete souls. The expectations of others are restraints; we must live for ourselves and not let others become burdens in our lives.

Figure 7: Random samples of answers chosen by CogVLM and MiniGPT-4, along with the standard answers, covering three categories: *Satirical*, *Critical*, and *Philosophical*, with one sample from each category.