Critical Review:
"Eyes Wide Shut? Exploring the Visual Shortcomings of Multimodal LLMs"

Summary

The paper "Eyes Wide Shut? Exploring the Visual Shortcomings of Multimodal LLMs" [1] investigates a significant challenge faced by multimodal large language models (MLLMs): they often fail to accurately detect low-level, straightforward visual details. The authors highlight a vulnerability where the reliance on vision encoders employed in MLLMs and that are pre-trained with contrastive language-image learning (e.g. CLIP) have a tendency to gloss over small but significant details and patterns. These include object count, orientation, colour and other such fine-grained features.

In Section 3 of "Eyes Wide Shut?", Tong et al. show how they implement the concept of "Clip-blind pairs" (see Figure 2), pairs of images that appear similar in CLIP's [2] embedding space yet are distinct in the reference self-supervised model (DINOv2). To address these issues, the authors introduce the Multimodal Visual Patterns (MMVP) benchmark. Designed to test MLLMs on simple questions regarding paired images, their empirical results infer that simply scaling CLIP on larger training sets or model sizes does not inherently resolve these vulnerabilities. Furthermore, their Table 1 outlines that open-source variants like LLaVA-1.5, also built on CLIP, fail to capture vital visual cues in advanced reasoning tasks.

Finally, Tong et al. address these issues by proposing an innovative method "Mixture-of-Features" that integrates a purely self-supervised vision encoder (DINOv2) with CLIP's language centric features. They go on to show that an additive fusion approach can sometimes degrade the model's ability; however, a more nuanced Interleaving tokens approach, named "interleaved-MoF" successfully preserves high-level textual alignment whilst simultaneously improving fine-grained detail recognition. In general, the paper highlights a latent vulnerability in MLLMs and concludes by offering practical directions for building more detail-sensitive architectures and systems.

Relation to Other Work

In the ever-growing research area of MLLMs, early approaches in visual question answering (VQA) tended to solely focus on datasets such as VQAv2 [3], which implemented multi-step reasoning but more often than not failed to isolate subtle visual discrepancies. Foundational work, such as the CLIP model [2], set the stage for contrastive language-image pre-training; however, its dependence on strict visual embeddings has led to limitations. Cutting-edge architectures such as MM-Vet [4] demonstrate that tasks requiring precise attribute recognition are not always feasible even in advanced systems. Li et al. [5] go beyond standard benchmarks by analysing and measuring how often a model perceives objects that are not actually present. This provides an interesting parallel to the "Eyes Wide Shut" demonstration of hallucinated details or missing crucial features.

Simultaneously, self-supervised learning in vision has seen rapid advancements spearheaded by architectures like DINOv2 and Masked Autoencoders (MAE) [6]. These methods can capture more fine-grained visual details that are often overlooked by more traditional CLIP-based encoders. In the same context, current models such as InstructBLIP demonstrate the potential ability of integrating robust visual features whilst keeping strong language understanding. However, Tong et al. [1] goes on to extend this research area by demonstrating how the CLIP architecture, despite its early successes in areas like zero-shot classification, crucially misinterprets details when asked relatively straightforward visual questions. This further imbues how "bridging methods" such as "Interleaved-Mof" proposed in "Eyes Wide Shut" could be paramount in improving MLLMs by combining visual grounding of self-supervised learning with text alignment capabilities of CLIP.

Strengths of the Paper

Tong et al. provide a systematic and controlled way of investigating detailed visual understanding through creating the MMVP benchmark from "CLIP-blind pairs". Systematic failures in misreading orientation or miscounting objects could now be identified, that would otherwise remain latent in broader tasks [1]. Moreover, one of the paper's most compelling strengths is its comprehensive empirical analysis. They test closed-source models like GPT-4V and other popular open-source models like LLaVA-1.5 on their benchmark and measure the

performance on human baselines. These results are clearly presented in Section 2.3 and Figure 3, wherein each model's shortfalls are demonstrated by specific question examples. Furthermore, when the models dip below random-guess accuracy on straightforward questions, the direct comparison with human responses highlights the importance of the identified weaknesses.

The innovative Mixture-of-Features approach, perhaps, is the most impactful contribution from the paper. By combining two cutting-edge systems, CLIP and a self-supervised vision model, DINOv2, and by interleaving their features they demonstrated measurable progress in visual grounding whilst keeping instruction-following capacity. Their results garner momentum for future MLLM optimisation, balancing improved visual understanding with textual alignment.

Weaknesses of the Paper

Despite its rigorous methodology, the paper ultimately leaves a few important gaps. First, while the authors demonstrate that missing crucial visual detail is not solved simply by making CLIP-based models larger in size or trained on enlarged datasets, they do not however thoroughly investigate alternative scaling methods. For instance, different training objectives or more specific fine-grained tasks might somewhat resolve these blind spots. Moreover, the Mixture-of-Features approach, while innovative and effective on the authors' benchmark, is left untested on interactive scenarios like dialogue-based instruction-following. Naturally, this leaves the reader to question its applicability on more complex tasks.

Furthermore, the dataset primarily draws on images from ImageNet and LAION, leaving ambiguity on whether CLIP-blind pairs are created similarly in more specialised domains like medical or satellite imagery. Recent research in region-based detection suggests local detail can greatly vary across domains, thus affecting performance [7]. Investigating these phenomena in more domain specific contexts could further legitimise the paper's claim that such oversights are universal. Finally, combining two large encoders (CLIP and DINOv2) can be expensive and there is relatively little discussion on potential mitigations, such as freezing layers or knowledge distillation that may reduce overhead. These issues may limit practical use in a resource-constrained setting and ultimately prohibit the deployment of an otherwise sophisticated approach.

Potential Advancement and Future Work

There are multiple avenues to explore that further build upon the current findings of "Eyes Wide Shut" by Tong et al. [1]. A promising such one is to extend the MMVP benchmark into more specialised domains, such as medical imaging or autonomous driving, where missed details could be critical. If CLIP-blind pairs exist in such settings, it would emphasise the necessity to re-calibrate visual encoders for more safety-oriented tasks. Another interesting avenue of research is to build Mixture-of-Features into a dynamic system. Similarly to how BLIP-2 [8] balances both text alignment with vision modules, the "Interleaved-MoF" model could gate self-supervised features when local detail is increasingly important. Potentially, this could reduce overhead of simultaneously running two large encoders, by combining fine-grained recognition with more efficient high-level reasoning and understanding.

Finally, a deeper investigation to look "under the hood" of CLIP's failure may discover more targeted and surgical fixes. For instance, incorporating local region objectives in the pre-training phase (e.g R-CNN) may perhaps introduce some sensitivity to counting or orientation from the start, as opposed to patching weaknesses later on. Moreover, future architectural model designs may be refined using interpretability methods like attention rollout, which may help identify where in the network smaller details are lost. To summarise, Tong et al. make a convincing argument in favour of a balanced integration of language-aligned and visual training that is crucial for a robust and detailed-sensitive MLLM.

References

[1] S. Tong et al, CVPR, 2024.

[2] A. Radford et al, ICML, 2021

[3] Y. Goyal et al, CVPR, 2017.

[4] W. Yu et al, arXiv, 2023.

[5] Y. Li et al, arXiv, 2023.

[6] K. He et al, CVPR, 2022.

[7] Y. Chen, CVPR, 2018.

[8] J. Li et al, ICML, 2023.