

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 [2]) 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 (Figure 2) how they implement the concept of “Clip-blind pairs”, pairs of images that appear similar in CLIP’s embedding space yet are distinct in the reference self-supervised model (DINOv2) [3]. To address these issues, the authors go on to 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 [4], also built on CLIP, fail to capture vital visual cues even when handling advanced reasoning tasks effectively.

Finally, Tong et al. address these issues by proposing an innovative method “Mixture-of-Features” that integrates a purely self-supervised vision encoder (DINOv2 [3]) 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 sophisticated and nuanced interleaving of tokens approach, named “interleaved-MoF” successfully preserves high-level textual alignment whilst simultaneously improving fine-grained detail recognition. Overall, the paper highlights a latent vulnerability in MLLMs and concludes by offering practical directions of 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 like VQAv2 [5] or similarly, Visual Genome [6], which implemented multi-step reasoning but more often than not failed to isolate subtle visual discrepancies. More cutting-edge architectures like Winoground [7] and MM-Vet [8] emphasised integrated and compositional capabilities, showcasing that tasks requiring precise attribute recognition is not always feasible even in advanced systems. Li et al. go beyond standard benchmarks in their paper “Evaluating object hallucination in large vision-language models” [9], 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 missed crucial features.

Self-supervised learning in vision has recently seen rapid advancements spearheaded by architectures like DINOv2 [3] and Masked Autoencoders (MAE) [10]. Tong et al. 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 embues how “bridging methods” such as “Interleaved-MoF”, could be paramount in improving and further driving the text driven nature of CLIP-based encoders.

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 be now identified, that would otherwise remain latent in broader tasks [1]. Furthermore, one of the paper’s most compelling strengths is its comprehensive empirical analysis. They test GPT-4V [11] and other popular open-source models like LLaVA-1.5 [4] on their benchmark and measuring 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. Moreover,

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 [2] and a self-supervised vision model, DINOv2 [3], 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 broader 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. Past research in region-based detection suggests local detail can greatly vary in domains where elements are partially occluded or repetitive [12]. Investigating these phenomenon 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

Multiple directions emerge to build upon the findings of "Eyes Wide Shut." One natural step is *extending MMVP* into more specialised or high-stakes areas, such as medical imaging, autonomous driving, or fine-grained species identification. If CLIP-blind pairs appear in such settings, it would underscore an urgent need to re-calibrate or retrain vision encoders for safety-critical tasks.

Further, the authors' "Interleaved-MoF" approach could evolve into a *dynamic gating* mechanism, wherein the model learns to emphasise the self-supervised features only when local detail is particularly important. Past frameworks like BLIP-2 have hinted at balancing text alignment with vision modules [?], and a gating-based approach could unify these insights for robust but flexible fusion.

An equally promising line of inquiry is *diagnosing the precise roots of CLIP's failure*. Perhaps introducing local region objectives—akin to early region-based detection networks [12, ?]—within the CLIP pipeline would highlight orientation or counting from the outset. Alternatively, training new vision-language models that explicitly integrate "pixel-level" self-supervision with textual alignment might preempt the need for post-hoc mixing. Finally, a deeper interpretability analysis, reminiscent of how large language model representations have been dissected [?], could reveal whether any architectural modules within CLIP systematically disregard small objects or positional cues.

In summary, "Eyes Wide Shut? Exploring the Visual Shortcomings of Multimodal LLMs" illuminates a significant gap in current vision-language paradigms: while global semantics are well captured, local details are not. By compiling a benchmark that exposes these blind spots and proposing a straightforward albeit computationally heavier fix, the authors make a compelling case that future MLLMs need more balanced, detail-oriented training. The paper's work thus stands as a timely reminder that continuing to escalate dataset size and model scale will not suffice to overcome fundamental representation issues. Instead, combining purely visual (self-supervised) and language-aligned approaches in a more struc-

tured, interpretability-conscious manner appears the most promising route to bridging the gap between robust local perception and fluent linguistic output.

References

- [1] S. Tong *et al.*, “Eyes wide shut? Exploring the visual shortcomings of multimodal LLMs” in *CVPR*, 2024.
- [2] A. Radford *et al.*, “Learning transferable visual models from natural language supervision” in *ICML*, 2021.
- [3] M. Oquab *et al.*, “DINOv2: learning robust visual features without supervision” *arXiv:2304.07193*, 2023.
- [4] H. Liu *et al.*, “Improved baselines with visual instruction tuning,” *arXiv:2310.03744*, 2023.
- [5] Y. Goyal *et al.*, “Making the V in VQA matter,” in *CVPR*, 2017.
- [6] R. Krishna *et al.*, “Visual genome: connecting language and vision using crowdsourced dense image annotations” *IJCV*, 2017.
- [7] T. Thrusch *et al.*, “Winoground: probing vision and language models for visio-linguistic compositionality” in *CVPR*, 2022.
- [8] W. Yu *et al.*, “MM-Vet: evaluating large multimodal models for integrated capabilities” *arXiv:2308.02490*, 2023.
- [9] Y. Li *et al.*, “Evaluating object hallucination in large vision-language models” *arXiv:2305.10355*, 2023.
- [10] K. He *et al.*, “Masked autoencoders are scalable vision learners,” in *CVPR*, 2022.
- [11] OpenAI, “GPT-4 technical report,” Tech. Rep., 2023.
- [12] R. Girshick *et al.*, “Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation,” in *CVPR*, 2014