This paper investigates the effects of vision representation in multimodal large language models (MLLMs). The authors specifically examine how MLLMs' performance is related to the cross-modal alignment and correspondence of the visual representation in various multimodal tasks.

The central research question is: "What fundamentally makes a feature representation achieve the highest performance?" To tackle this, they introduce the concept of cross-modal alignment and visual correspondence (AC score), which quantifies the relationship between vision encoding methods and model outcomes.

One major challenge the authors encountered was the high computational cost

in finding the optimal visual representations, which is due to the need of finetuning the language model under each setting. Authors claim that by following the proposed AC policy, one can reasonably predict the MLLM's performance given different combinations of visual representations, enhancing both accuracy and efficiency compared to randomly searching for the optimal representation.

Their claim is supported by experimental results involving thirteen distinct vision representation settings across eight benchmarks. Under their definition, the AC score and model performance exhibit a linear relationship, with a coefficient of determination of 95.72%. Additionally, the authors compare the effectiveness of using AC scores against random or single-factor scores, further validating their methodology.

Overall, the paper proposes that integrating AC scores into the vision representation selection process could greatly enhance both efficiency and effectiveness in developing robust multimodal AI systems.