> Q4: While the authors claim at line 499 that MEMVR is flexible and compatible with various architectures, they also mention at line 514 that adapting the method to other structures is challenging. It raises the question of whether MEMVR can be generalized to models like LLaVA-next and InstructBlip.

First, sorry for the confusing words above. In line 499, we regard MEMVR as a flexible approach as it only take few lines of code without modifying transformers lib, and therefore can be easily applied to other VLMs. In line 514, where we discussed about the limitations of our work, we aim to express that MEMVR needs few rounds of adjustments to achieve optimal performance when applying to a brand-new model due to the difference of optimal hyperparameter across different models, and the manually tuning procedure makes it relatively challenging to run with optimal performance—as the tuning procedure takes time. We take it as kind of trade-off for having best performances in most track in all 9 benchmark evaluations.

Secondly, we’re confident to say that MEMVR can be easily generalized to other models like LLaVA-Next, LLaVA1.6-mistral or others. But we’d like to discuss with you about a certain case: InstructBlip, which is the only model that does not work well with MEMVR. InstructBlip uses a very special Q-Former that transforms visual information into few tokens (often no more than 20 tokens). This makes the MEMVR calculation very difficult to keep the retraced information and conduct attention-like operation, as the core calculation of MEMVR can be shorten as SiLU(hidden\_state\*visual\_token.T)\*visual\_token. Such short visual token brought by InstructBlip will lead to the fact that the retraced information will be compressed into a (t,t) matrix where t is usually less than 20, while the original visual information is represented as (t, depth) where the depth is much higher than t. Therefore conducting MEMVR calculation with InstructBlip does not seem to have any effect on improving the quality of generated answers.

In all, IntructBlip is just a very special case across several models we’ve tested, it’s poor performance with MEMVR is due to its Q-Former architecture that could lead to information loss. MEMVR can still be easily applied to almost all open-source VLMs with significant improvements in both hallucination mitigation and general multi-modal understanding.