Reducing Hallucinations in Vision-Language Models via Latent Space Steering

Sheng Liu Haotian Ye Lei Xing James Zou

Stanford University

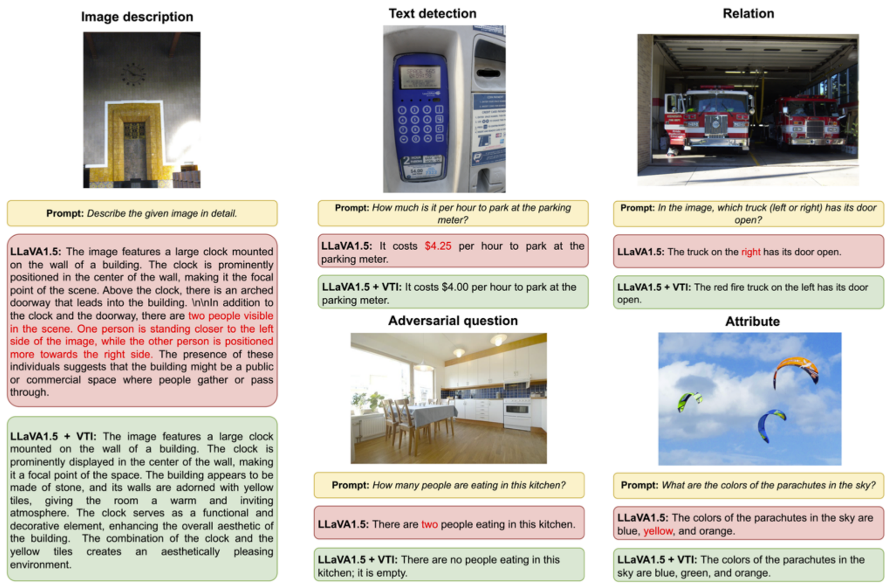

Figure 1: Illustration of the effect of our proposed method, VTI, using LLaVA-1.5. Hallucinated contents generated by the original model are marked in red. In contrast, VTI results in less hallucination across different categories of questions. Examples are obtained from MMHAL-Bench (Sun et al., 2023) and CHAIR (Rohrbach et al., 2018)

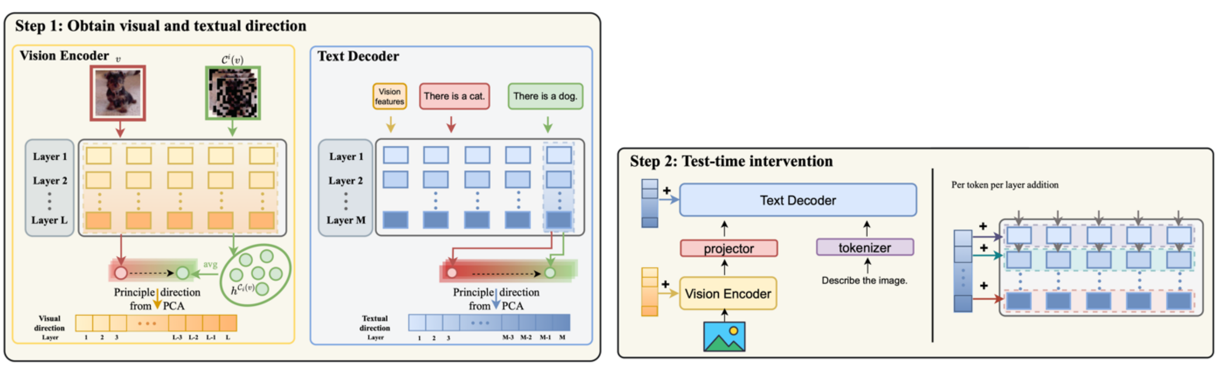

Figure 3: Overview of the proposed algorithm visual and textual test-time intervention (VTI). Given an example set $\{(v_i, x_i, \tilde{x}_i)\}_{i=1}^N$ where v_i is the vision input and $(x_i, \tilde{x}_i)$ is paired captions with and without hallucination, VTI first runs the model on each query $(v_i, x_i, \tilde{x}_i)$ and records all hidden states. It then computes the shifting vectors $d_{l,t}^{\text{vision}}$ and $d_{l,t}^{\text{text}}$ for all layer l and token t according to Section 4. During inference, the vectors are subsequently added to every layer of the vision encoder and text decoder, respectively, when processing a new query. Notice that the vectors are task- and dataset-agnostic, i.e., they are pre-computed using a few samples from one specific task and dataset, and fixed unchanged throughout the entire experiments in our paper.