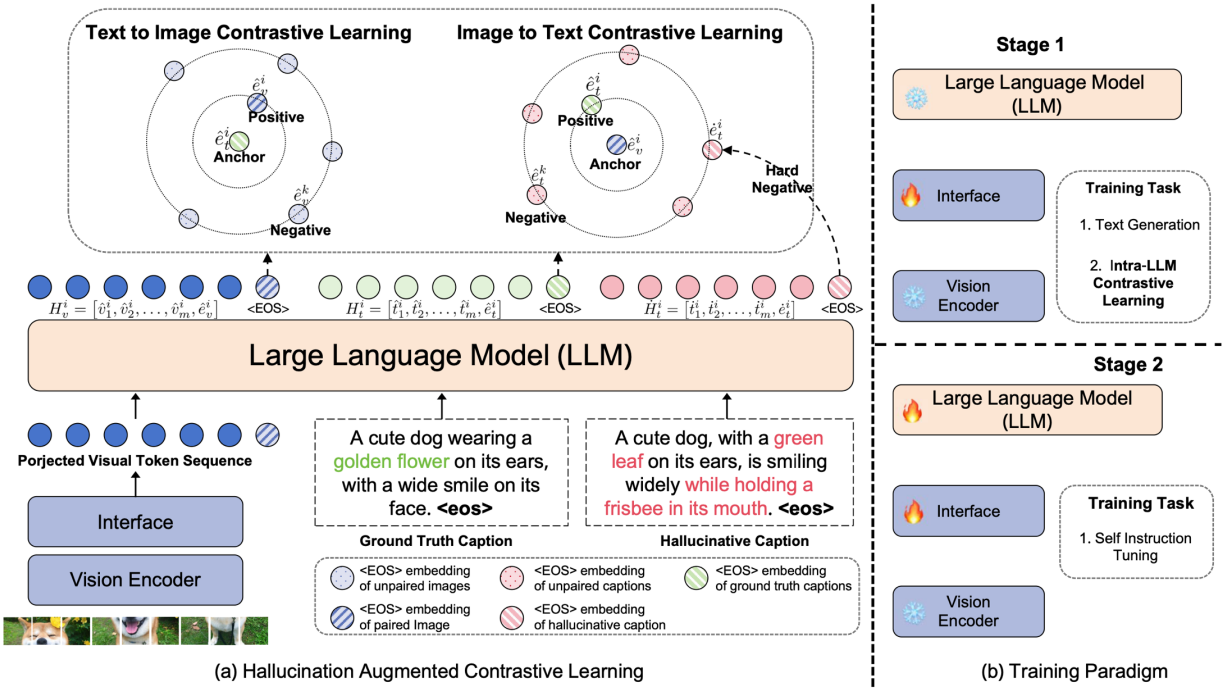
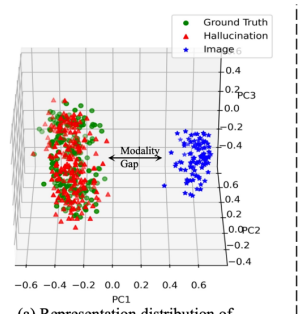
Hallucination Augmented Contrastive Learning for Multimodal Large Language Model

Chaoya Jiang[1] Haiyang Xu[2*] Mengfan Dong[1] Jiaxing Chen[1] Wei Ye[1*]
Ming Yan[2] Qinghao Ye[2] Ji Zhang[2] Fei Huang[2] Shikun Zhang[1]
[1]National Engineering Research Center for Software Engineering, Peking University
[2]Alibaba Group
{jiangchaoya, wye}@pku.edu.cn, shuofeng.xhy@alibaba-inc.com

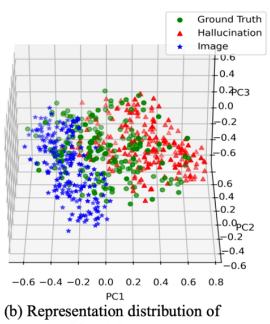


(a) Hallucination Augmented Contrastive Learning (b) Training Paradigm

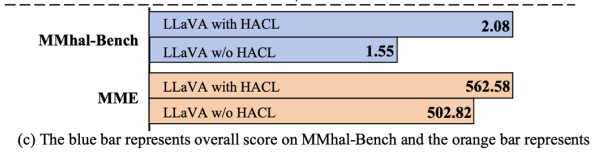
Figure 2. Subfigure (a) illustrates the proposed HACL. In this framework, we employ GPT-4 [38] to generate the hallucinative captions as the the hard negative samples in the image-to-text contrastive learning. Subfigure (b) shows the training paradigm of HACL.



(a) Representation distribution of LLaVA w/o HACL



(b) Representation distribution of LLaVA with HACL



(c) The blue bar represents overall score on MMhal-Bench and the orange bar represents the overall score on MME. Noted biger numbers mean better results.

Figure 3. This figure showcases a range of hallucinative captions generated by GPT-4. The hallucinative text is highlighted in red.