ICT: Image-Object Cross-Level Trusted Intervention for Mitigating Object Hallucination in Large Vision-Language Models

Supplementary Material

Appendix A: Implementation Details

In our experiments, we used greedy search to ensure reproducibility. Hyperparameter tuning was performed using a grid search to explore the possible combinations of key parameters systematically. The hyperparameters under consideration included α, β, and K. To reduce the dimensionality of the search space, we constrained α and β to be equal. This constraint is reasonable as these two hyperparameters serve similar roles in our experiments. For LLaVA-v1.5, we applied interventions to the layers `language_model.model.layers.{i}.self_attn.o_proj`, while for Qwen-VL, the interventions were applied to `transformer.h.{i}.attn.c_proj`.

Grid Search Process The grid search exhaustively evaluated all combinations of hyperparameters within the specified ranges. Given the aforementioned constraint on α and β, the search space was defined as the Cartesian product:

$$\{(8,8),(16,16),(24,24),(32,32)\} \times \{32,64,128,256\}.$$

This resulted in a total of $4 \times 4 = 16$ unique hyperparameter configurations, which is a relatively small search space.

Hyperparameter Search Details

α and β: The hyperparameters α and β, which control the strength of Image-Level and Object-Level intervention, respectively, were allowed to take values from the discrete set $\{8, 16, 24, 32\}$. This range was chosen to strike a balance between improving model trustworthiness and maintaining overall performance. By enforcing the constraint $\alpha = \beta$, the number of unique combinations was reduced to four, thereby simplifying the search space while ensuring sufficient exploration of the parameter landscape.

K: The hyperparameter K, representing the number of attention heads we intervene on, was explored over the set $\{32, 64, 128, 256\}$. This range was determined based on two key considerations:

- Both LLaVA-v1.5 and Qwen-VL architectures consist of 32 layers, each with 32 attention heads, resulting in 1024 total attention heads. The chosen range allows for intervention on varying subsets of these attention heads.
- Preliminary classification accuracy results, as shown in Figure 7, indicate that this range provides sufficient flexibility to cover scenarios from intervening on all heads that are relevant to hallucination to intervening only on those most pertinent to truthfulness.

Methods	**PhD Attribute**	**PhD Positional**	**MMMU**
Regular-7B	59.44	61.05	33.1
VCD-7B	62.45	63.01	33.4
ICT-7B	66.87**(+7.43)**	69.18**(+8.13)**	34.8**(+1.7)**
Regular-13B	63.29	60.89	36.4
VCD-13B	71.84	69.71	34.6
ICT-13B	74.63**(+11.34)**	72.05**(+11.36)**	37.7**(+1.3)**

Table 4. Performances on PhD-base and MMMU using different methods based on LLaVA-v1.5.

Appendix B: More Case Studies

In Figure 8, we present additional illustrative cases from LLaVA-bench to further demonstrate the effectiveness of ICT on extremely challenging open-ended questions. The upper part of Figure 8 vividly illustrates the precision and reliability of our intervention. ICT retains most of the original phrasing, selectively eliminating segments containing untruthful content. This demonstrates that ICT not only significantly reduces the model's tendency to hallucinate but also achieves this with minimal and minimal side effects.

The lower part of Figure 8 presents another example where ICT successfully preserves and utilizes useful language priors, whereas VCD fails to do so. In this case, both approaches accurately identified the building, but VCD hallucinated incorrect scene opening hours. In contrast, ICT provided suggestions grounded in factual knowledge, leveraging the positive aspects of language priors effectively.

Appendix C: More details on Probe Training

To train the binary classifier, we use the activation value vectors of the original image/text pairs as positive samples and those of global/local blurry images and text pairs as negative samples. This allows us to train a binary classifier, we use SVM throughout this paper. We hypothesize that if an attention head encodes relevant information, the activation vectors produced by that head should show significant differences, leading to high classification accuracy. In practice, we select the top-K heads with the highest accuracy across all heads as the corresponding image heads or object heads. This approach enables us to identify the heads encoding relevant information in a fine-grained and adaptive manner, without needing to manually select specific layers for intervention.

Appendix D: Additional Experiments of ICT

Generalizability and Model Selection To address possible concerns regarding whether ICT could generaize be-

Figure 8. More case studies of ICT. In the first example, our ICT avoids captioning non-existent objects in the image while keeping the rest of the caption nearly identical. This demonstrates its ability to refine the accuracy of generated content while maintaining the quality and coherence of the text. In the second example, VCD hallucinated the opening hours of the Space Needle, whereas our ICT correctly retrieved this information from its internal knowledge, stored implicitly within the language modality. This highlights our method's remarkable capability to minimize vision-related hallucinations while leveraging the advantages of language priors.

yond object existence tasks and whether we conducted additional experiments focusing on relation and attribute hallucinations, as well as broader reasoning capabilities. Specifically, we expanded our evaluation to include PhD-base (Attribute and Positional subsets) [57] and MMMU [99] , which emphasize multimodal reasoning. Additionally, we increased the diversity of models by incorporating the larger LLaVA-v1.5-13B to assess the impact of activation modifications across different model scales. We applied the COCO Random activation shift vectors to the model and tested on the PhD-base benchmark without additional modifications for different types of hallucinations. Table 4 shows that ICT still greatly improved performance, further demonstrating its robustness. We have included results from using LLaVA-v1.5-7B and LLaVA-v1.5-13B on the MMMU benchmark, using the COCO Random activation shift vectors as well. As shown in Table 4, unlike contrastive decoding methods such as VCD, which eliminate language priors to mitigate hallucinations—potentially removing useful language priors for reasoning—our method enhances the model's attention to visual information through intervention transfer vectors. This approach not only addresses the issue of hallucinations but also improves performance on general benchmarks, including MMMU. The above-mentioned results further prove that our ICT generalize well across different models of different sizes on various types of tasks.

Setting	**Methods**	Accuracy	F1 Score
Random	ICT-7B-200	89.83	89.44
	ICT-7B-1500	90.11	90.03
	Regular-13B	83.31	81.49
	VCD-13B	87.39	86.55
	ICT-13B-1500	90.67**(+7.36)**	90.33**(+8.84)**
Popular	ICT-7B-200	85.80	86.71
	ICT-7B-1500	87.50	87.60
	Regular-13B	82.47	80.75
	VCD-13B	85.74	85.06
	ICT-13B-1500	87.86**(+5.39)**	87.74**(+6.99)**
Adversarial	ICT-7B-200	84.00	83.15
	ICT-7B-1500	84.43	83.74
	Regular-13B	80.00	78.62
	VCD-13B	81.92	81.78
	ICT-13B-1500	85.01**(+5.01)**	84.75**(+6.13)**

Table 5. Performance of different methods based on LLaVA-v1.5 on the POPE COCO dataset. The numbers represent the sample size used by ICT.

Impact of Dataset Size Furthermore, activation shift vectors acquisition is not sensitive to the size of the dataset: the activation shift vectors of 200 COCO entries showed similar performance on COCO subsets compared to 1,500 entries, as shown in Table 5. Using shifting vectors obtained from 200 samples yields similar results to to those of of 1500 samples, indicating that the size of the dataset does not significantly influence the performance of our ICT.