TruthPrInt: Mitigating Large Vision-Language Models Object Hallucination Via Latent Truthful-Guided Pre-Intervention

Supplementary Material

A. Hallucination Detection with Internal States

A.1. Internal States Collection

In Sec. 3.1, we utilize the hidden states of preceding tokens associated with object tokens to detect hallucinations. Specifically, the hallucination detector is designed to provide an early warning by predicting whether future object tokens are likely to be hallucinated. This approach ensures that the detector is not exclusively trained on object tokens but functions as a generalized detector applicable to any type of token. From an intervention perspective, this "early warning" mechanism reduces the inference time of the LLM during decoding. For example, when determining the next token z_j, the previous hidden states can be directly passed to the detector for hallucination identification, i.e., $\mathcal{G}(\boldsymbol{h}_{j-1}) < \tau$. In contrast, a "current-token" prediction approach would require computing the current hidden states $\boldsymbol{h}_j$, which involves an additional LLM inference step before detecting hallucinations, i.e., $\mathcal{G}(\boldsymbol{h}_j) < \tau$.

A.2. Training Protocol of Hallucination Detection

In our implementation, the hallucination detector $\mathcal{G}$ is a 3-layer MLP, with the architecture presented in Tab. 6. The model is trained for 30 epochs with a batch size of 512, a learning rate of 0.001, and the Adam optimizer, utilizing binary cross-entropy (BCE) as the training objective. The optimal checkpoint is determined based on its performance on the validation set.

Layer 1	Layer 2	Layer 3	Activation
(4096, 128)	(128, 64)	(64, 1)	ReLu

Table 6. The architecture of $\mathcal{G}$.

B. TruthPrInt: Preliminary Analysis

B.1. Low-Confidence Tokens Precede Hallucination

As we mentioned in Sec. 4.2, tokens with lower confidence frequently precede hallucinated objects. Here, we provide experimental evidence to support it. Specifically, for each object token, we calculate *Preceding Minimum Confidence* (**PMC**): the minimum LVLM confidence of the preceding tokens of the object token within the same sentence. In Tab. 7, we present the average PMC collected from hallucinated object tokens and truthful object tokens, respectively, over 500 samples. It is shown that the PMC of hallucinated is significantly larger than the PMC of truthful object tokens, indicating that low-confidence tokens tend to derive hallucinated objects.

Model	PMC of Hallucinated	PMC of Truthful
MiniGPT-4	0.39	0.31
Llava-1.5	0.29	0.22
mPlug-Owl2	0.29	0.20

Table 7. The average *Preceding Minimum Confidence* (PMC) over hallucinated and truthful object tokens. The PMC of hallucinated objects is significantly larger than the PMC of truthful object tokens, indicating that tokens with lower confidence frequently preceded hallucinated objects.

B.2. Method Procedures

In Algorithm 1, we present our pre-intervention mechanism algorithmic descriptions.

C. Experiment Protocols

In this section, we introduce the OH benchmarks used in this paper and additional experimental results as well.

C.1. Benchmarks

MSCOCO CHAIR [43] is a widely used benchmark for evaluating OH. Given a set of images, it tasks LVLMs with generating detailed descriptions of the images. The next step involves comparing the objects present in the images with those mentioned by the LVLMs, using specific metrics

$$\text{CHAIR}_S = \frac{|\text{sentences with hallucinated objects}|}{|\text{all sentences}|}$$

$$\text{CHAIR}_I = \frac{|\text{hallucinated objects}|}{|\text{all objects mentioned}|}$$

for OH evaluation. It is usually incorporated with the COCO image caption dataset.

POPE [28] conducts an empirical evaluation of OH across multiple LVLMs, revealing its severity and identifying critical factors influencing this issue. It introduces Polling-based Object Probing Evaluation (POPE), which reformulates hallucination assessment as a binary classification task to improve stability, fairness, and scalability over existing methods.

LLaVA-Bench [31] is a diverse collection of 24 images featuring various contexts, such as in-door, and outdoor. Each

Algorithm 1 `TruthPrInt` decoding

Input: Prompt $\boldsymbol{s}$, model $\mathcal{M}$, the image $\boldsymbol{x}$, max backtracing number $\mathcal{N}_B$, detector $\mathcal{G}$, target layer L, threshold τ
$k = 0$, $i = 0$
$\boldsymbol{r} = \mathbf{0}$ ▷ Rank of Selected Token
$\boldsymbol{c} = \mathbf{0} \in \mathbb{N}^{\mathcal{N}_B+1}$ ▷ # of Hallucination
repeat
  repeat ▷ Generate a Sentence
    $\boldsymbol{o}_i^k = \mathcal{M}^o(\boldsymbol{x}, \boldsymbol{s}, \boldsymbol{z}_{<i}^k; \boldsymbol{\theta})$
    $\boldsymbol{h}_{i-1}^k = \mathcal{M}^L(\boldsymbol{x}, \boldsymbol{s}, \boldsymbol{z}_{<i}^k; \boldsymbol{\theta})$
    $z_i^k = \text{TopK}(\boldsymbol{o}_i^k, \boldsymbol{r}_i + 1)$ ▷ Next Rank Token
    $\boldsymbol{c}_k = \boldsymbol{c}_k + \mathbb{1}[\mathcal{G}(\boldsymbol{h}_i^k) > \tau]$
    $\boldsymbol{r}_i = \boldsymbol{r}_i + \mathbb{1}[\mathcal{G}(\boldsymbol{h}_i^k) > \tau]$
    $i = i + 1$ ▷ Generate Next Token
  until z_{i-1}^k in $[eos, .]$
  if $\boldsymbol{c}_i = 0$ **then** ▷ No Hallucination
    return $\boldsymbol{z}^k$
  else ▷ Next Backtracing Initialization
    $k = k + 1$
    $i^k = \arg\min(\{\text{TopK}(\boldsymbol{o}_j^{k-1}, 1) | j \leq i\})$
    $\boldsymbol{z}_{<i^k}^k = \boldsymbol{z}_{<i^k}^{k-1}$, $i = i^k$
    $\boldsymbol{r}_{>i} = 0$ ▷ Set State and Backtracing From i^k
  end if
until $k > \mathcal{N}_B$ ▷ Achieve the Max Backtracing Number ▷ Find Sentence with Less Hallucination
$k' = \arg\min(\boldsymbol{c}_{\leq \mathcal{N}_B})$
$i = \text{FindFirstHallucination}(\boldsymbol{z}^{k'})$
$\boldsymbol{z}_{<i}^k = \boldsymbol{z}_{<i}^{k'}$ ▷ Backtracing from i
repeat
  $\boldsymbol{o}_i^k = \mathcal{M}^o(\boldsymbol{x}, \boldsymbol{s}, \boldsymbol{z}_{<i}^k; \boldsymbol{\theta})$
  $\boldsymbol{h}_{i-1}^k = \mathcal{M}^L(\boldsymbol{x}, \boldsymbol{s}, \boldsymbol{z}_{<i}^k; \boldsymbol{\theta})$
  $z_i^k = \text{TopK}(\boldsymbol{o}_i^k, \mathbb{1}[\mathcal{G}(\boldsymbol{h}_{i-1}^k) > \tau] + 1)$
  $\boldsymbol{c}_k = \boldsymbol{c}_k + \mathbb{1}[\mathcal{G}(\boldsymbol{h}_{i-1}^k) > \tau]$
  $i = i + 1$
until z_{i-1}^k in $[eos, .]$
$k = \arg\min(\boldsymbol{c})$
return $\boldsymbol{z}^k$

image is paired with a meticulously crafted, detailed description and a thoughtfully chosen set of questions. It is usually used for quantitative analysis of LVLM behaviors.

C.2. POPE Results

In Tab. 9, we present the individual results over each offline POPE split. We also provide the original POPE evaluation results, obtained from MiniGPT-4 for each split, in Tab. 8.

C.3. LLaVA-Benchmark Quantitative Analysis

We evaluate our methods and baselines on the LLaVA-Benchmark (In-the-Wild) dataset, manually reviewing the generated responses for these images (Fig. 9). Our observations reveal that `TruthPrInt` produces more accurate and truthful descriptions, with greater detail included compared to the baselines.

Method	Random		Popular		Adversarial		*average*	
	Precision↑	F_β ↑	**Precision**↑	F_β ↑	**Precision**↑	F_β ↑	**Precision**↑	F_β ↑
Greedy	67.65	67.78	55.60	55.79	58.97	59.15	60.74	60.91
VCD	60.76	60.79	52.63	52.70	54.33	54.38	55.91	55.96
Beam	64.30	64.47	54.68	54.88	56.44	56.64	58.47	58.66
TruthPrInt	**68.23**	**68.35**	**55.76**	**55.93**	**59.09**	**59.26**	**61.03**	**61.18**

Table 8. Evaluation results on the original POPE benchamrk.

POPE Split	Methods	MiniGPT4		Llava-1.5		mPlug-Owl2	
		Precision↑	F_β ↑	**Precision**↑	F_β ↑	**Precision**↑	F_β ↑
Random	Greedy	97.13±0.22	95.59±0.16	98.21±0.16	**96.95±0.06**	96.66±1.44	95.39±1.46
	Beam	97.51±0.92	95.93±0.80	97.70±0.14	96.43±0.22	96.47±1.76	95.05±1.67
	VCD	96.78±1.42	95.14±1.35	97.11±1.18	95.91±1.06	96.80±0.87	**95.40±0.84**
	OPERA	98.12±0.51	96.51±0.44	97.70±0.46	96.43±0.48	96.10±1.30	94.62±1.15
	DOLA	97.51±0.52	**95.94±0.46**	97.70±0.12	96.43±0.17	96.47±1.35	95.04±1.27
	HALC	97.04±0.39	95.33±0.38	97.98±1.01	96.60±1.00	96.73±1.24	95.35±1.20
	TruthPrInt	**98.17±0.46**	95.58±0.46	**98.65±0.80**	96.63±0.86	**97.48±0.64**	95.28±0.71
Popular	Greedy	87.50±2.16	86.34±2.10	91.63±1.32	90.60±1.36	89.69±1.36	88.66±1.26
	Beam	89.61±1.01	88.34±1.04	90.92±0.50	89.88±0.41	90.30±3.05	89.12±2.97
	VCD	87.12±0.87	85.87±0.74	91.11±1.69	90.11±1.66	89.18±0.46	88.07±0.40
	OPERA	88.85±0.84	87.61±0.85	90.52±2.19	89.49±2.14	89.42±1.22	88.21±1.26
	DOLA	90.13±0.19	**88.85±0.22**	91.14±0.25	90.09±0.19	90.01±2.72	88.83±2.65
	HALC	89.16±1.51	87.79±1.44	90.90±1.10	89.86±1.10	89.50±1.10	88.39±1.06
	TruthPrInt	**90.23±1.66**	88.10±1.47	**93.13±0.86**	**91.38±0.90**	**92.64±1.75**	**90.70±1.76**
Adversarial	Greedy	85.75±1.53	84.64±1.48	88.56±2.07	87.63±2.08	87.82±1.79	86.85±1.71
	Beam	87.59±0.22	86.40±0.28	90.69±0.83	**89.66±0.73**	89.58±0.40	88.42±0.41
	VCD	85.64±1.53	84.45±1.38	88.78±1.97	87.85±1.92	86.24±0.61	85.23±0.58
	OPERA	86.97±0.80	85.80±0.73	89.78±0.54	88.77±0.55	88.07±1.33	86.91±1.36
	DOLA	88.10±0.60	**86.90±0.54**	90.56±1.07	89.53±0.99	89.29±0.26	88.13±0.22
	HALC	87.44±2.65	86.13±2.54	88.99±1.17	88.02±1.17	87.55±1.59	86.50±1.55
	TruthPrInt	**88.44±2.09**	86.42±2.05	**91.06±1.19**	89.40±1.05	**90.86±1.82**	**89.01±1.82**

Table 9. Evaluation results of each offline POPE split.

Greedy: This image shows a person holding a smartphone with a charging cable attached to it. The phone is on a table with a white background. The person is wearing a white shirt and black pants. The charging cable is blue and has a small plug on the end. The phone is turned on and has a screen that shows a message. The person is holding the phone with their left hand and the charging cable with their right hand. The image is taken in a room with a white wall and a brown floor. The image shows a person holding a smartphone with a charging cable attached to it.

HALC: This image shows a person holding a smartphone with a charging cable attached to it. The phone is on a couch with a white blanket on it. There is a small table next to the couch with a white cup on it. The person is wearing a white shirt and blue The image shows a person holding a smartphone with a charging cable attached to it. The phone is on a couch with a white blanket on it. There is a small table next to the couch with a white cup on it. The The image shows a person

Broken Sentence

TruthPrInt: The image shows a person holding a smartphone with a charging cable attached to it. The phone has a small screen on the front. The charging cable is white and has a small plug on the end. There is a small amount of dust on the screen. The image is taken in a room with a wooden floor and a white wall. There is a window on the left side of the image and a door on the right side.

(a) Both Greedy and HALC hallucinate details such as a person wearing a **white shirt and blue**, along with other nonexistent objects like a "*phone message*" and a "*couch*". In contrast, `TruthPrInt` delivers more accurate and truthful descriptions.

Greedy: The image shows a group of purple mangos on a wooden table. The fruit is cut open and the inside is visible. There are several slices of the fruit on the table. The background is a wooden floor with a green plant in the corner.

HALC: The image shows a group of purple mangos on a wooden table. The fruit is cut open and the inside is visible. There are several slices of the fruit on the table. The background is a wooden surface with a green plant in the The image is in focus and well lit.

Broken Sentence

TruthPrInt: The image shows a group of purple fruit, including one that has been cut open to reveal the inside. The fruit is surrounded by a wicker basket and there are several other fruits in the background. The overall mood of the image is healthy and fresh.

(b) Both Greedy and HALC incorrectly describe the item as a **purple mango** and further hallucinate details like "*several slices of this fruit*". In contrast, `TruthPrInt` offers a more accurate description, referring to it as **purple fruits**.

Greedy: The image shows a group of animals sitting around a table with a tea pot and cups. The animals are a bear, a cat, and a rabbit. They are all wearing hats and are enjoying their tea together. The background is a green forest with trees and bushes. The image is watercolor and has a soft, dreamy quality to it. The colors are muted and pastel, with a focus on earth tones. The animals are drawn in a simple, cartoonish style, with exaggerated features and expressive faces. The overall mood of the image is cozy

HALC: The image shows a group of animals sitting around a table with a tea set and a plate of food. The animals are a bear, a cat, and a dog. They are all wearing hats and are enjoying their meal together. The background is a green forest with trees and bushes. The image is colored and has a soft, dreamy quality to it. The animals are drawn in a simple, cartoonish style. The overall atmosphere of the image is peaceful and serene.

TruthPrInt: The image is a watercolor painting of a bear, a cat, and a rabbit sitting at a picnic blanket in a forest. The bear is holding a plate of food, the cat is sitting on the plate, and the rabbit is sitting on the ground next to the plate. The background is a green forest with trees and bushes. The painting is done in a realistic style with watercolor paints. The colors used are earth tones such as brown, green, and beige.

(c) Both Greedy and HALC falsely describe all the animals as **wearing hats** and provide only limited details about the image. Additionally, **HALC misidentifies the rabbit as a dog**. In contrast, `TruthPrInt` delivers accurate descriptions of all the animals and includes additional details such as "*the bear is holding a plate of food*" and "*the colors used are earth tones like brown, green, and beige*".

Figure 9. LLaVA-Bench quantitative analysis results.