Stop learning it all to mitigate visual hallucination, Focus on the hallucination target.

Supplementary Material

A. Side Effects of Preference Learning

Here, we observe changes in the attention map to evaluate the side-effect of preference learning on mitigating hallucinations. Details in figure 5.

B. Proof of Theory

Lemma 1. *Given a reward function $r(x, y)$, assuming Assumption 3.1, we can establish the equivalence between the Bradley-Terry model to the 2.2.*

Proof. We need to show that under Assumption 1, the traditional Bradley-Terry model applied to the full response is equivalent to Equation 2.2 (our assumed equation) applied to the target chunks. 1. Traditional Bradley-Terry Model (Applied to Full Response)

$$P(y_r \succ y_h \mid x) = \sigma(r(x, y_r) - r(x, y_h))$$

Where, y_r is the revised full response, $y_h l$ is the hallucinated full response, $r(x, y)$ is the reward for the full response y given input x.

We are given that $\sum r_t = \sum r$. In our case, this means:

$$r(x, y_r) = r_t(x, y_r^t), r(x, y_h) = r_t(x, y_h^t)$$

This is because Assumption 3.1 states that the reward difference is entirely contained within the target chunks.

Substituting the values from step 2 into the traditional Bradley-Terry model, we get:

$$P(y_r \succ y_h \mid x) = \sigma(r_t(x, y_r^t) - r_t(x, y_h^t))$$

This is exactly the same as Equation 2.2 (our assumed equation):

$$P(y_r^t \succ y_h^t \mid x) = \sigma(r_t(x, y_r^t) - r_t(x, y_h^t))$$

□

Theorem 1. *Under Assumption 3.1, preference learning remains equivalent when non-target tokens are excluded.*

Proof. We need to show that:

$$\mathbb{E}y \sim \pi_\theta[r(x, y)] = \mathbb{E}y^t \sim \pi_\theta[r(x, y^t)]$$

Left-hand side (Original RLHF): $\mathbb{E}y \sim \pi_\theta[r(x, y)]$ This represents the expected reward over all possible responses y generated by the policy π_θ given the input x.

Right-hand side (Target-restricted RLHF): $\mathbb{E}y^t \sim \pi_\theta[r(x, y^t)]$ This represents the expected reward over all possible target chunks y^t generated by the policy π_θ given the input x.

Assumption 3.1 states that $\sum r_t = \sum r$. In the context of expected values, this implies:

$$r(x, y) = r(x, y^t)$$

for any response y and its corresponding target chunks y^t. Therefore:

$$\mathbb{E}y \sim \pi_\theta[r(x, y)] = \mathbb{E}y \sim \pi_\theta[r(x, y^t)]$$

Since the reward function only considers the target chunks y^t due to Assumption 3.1, we can replace the expectation over all responses y with the expectation over target chunks y^t without changing the value:

$$\mathbb{E}y \sim \pi_\theta[r(x, y^t)] = \mathbb{E}y^t \sim \pi_\theta[r(x, y^t)]$$

Thus:

$$\mathbb{E}y \sim \pi_\theta[r(x, y)] = \mathbb{E}y^t \sim \pi_\theta[r(x, y^t)]$$

This proves the equivalence of the expected rewards.

We also need to show that the policy gradient update for target-restricted RLHF follows the same form as for the original RLHF.

Original RLHF Gradient Update:

$$\nabla_\theta \mathbb{E}_{y \sim \pi_\theta}\left[r(x, y) - \beta D_{\text{KL}}(\pi_\theta | \pi_{\text{ref}})\right]$$

Target-restricted RLHF Gradient Update:

$$\nabla_\theta \mathbb{E}_{y^t \sim \pi_\theta}\left[r(x, y^t) - \beta D_{\text{KL}}(\pi_\theta | \pi_{\text{ref}})\right]$$

Since we have already established that $\mathbb{E}y \sim \pi_\theta[r(x, y)] = \mathbb{E}y^t \sim \pi_\theta[r(x, y^t)]$, we can substitute this into the original RLHF gradient update:

$$\begin{aligned}\nabla_\theta \mathbb{E}_{y \sim \pi_\theta}[r(x, y) - \beta D_{\text{KL}}(\pi_\theta | \pi_{\text{ref}})] = \\ \nabla_\theta \mathbb{E}_{y^t \sim \pi_\theta}[r(x, y^t) - \beta D_{\text{KL}}(\pi_\theta | \pi_{\text{ref}})]\end{aligned} \tag{7.1}$$

This shows that the policy gradient updates for the original RLHF and target-restricted RLHF are the same under Assumption 3.1.

As previously proven, under Assumption 3.1, since $r(x, y) = r(x, y^t)$, we can replace each response pair (y_r, y_h) in the preference dataset $\mathcal{D}$ with the corresponding target chunks (y_r^t, y_h^t). Therefore, the objective function of the target-learning DPO is as follows:

$$L_{\text{TL-DPO}}(\theta) = - \mathbb{E}_{(x,y_r^t,y_h^t)\sim D}\left[\log \sigma\left(\beta \log \frac{\pi_\theta(y_r^t|x)}{\pi_{\text{ref}}(y_r^t|x)} - \beta \log \frac{\pi_\theta(y_h^t|x)}{\pi_{\text{ref}}(y_h^t|x)}\right)\right] \tag{7.2}$$

In conclusion, we can see that the existing preference learning method including RLHF, DPO can be applied in the same way to target learning.

□

Proposition 1. ***Efficiency comparison in target learning*** *Let $\mathcal{H}_{pl}$ and $\mathcal{H}_{tl}$ be the hypothesis spaces to learning methods without target(pl) and with target(tl), respectively. The number of samples required to achieve the same generalization error ϵ and confidence level $1 - \delta$ satisfies $m_{tl} \leq m_{pl}$.*

Proof. Target learning restricts the problem by focusing on a subset y^t of the full output y. The functions in $\mathcal{H}_{tl}$ only need to discriminate based on variations within y^t. In contrast, functions in $\mathcal{H}_{pl}$ must accommodate variations across the entire y. Since y^t represents a smaller, more specific part of the output space compared to y, the class of functions needed to model preferences over y^t ($\mathcal{H}_{tl}$) is inherently less complex than the class needed for y ($\mathcal{H}_{pl}$). While any preference function in $\mathcal{H}_{tl}$ can be represented within $\mathcal{H}_{pl}$ (by ignoring non-target parts), $\mathcal{H}_{pl}$ must also contain functions sensitive to variations outside y^t, which are explicitly excluded from consideration in $\mathcal{H}_{tl}$. Therefore, the complexity of $\mathcal{H}_{tl}$ is strictly less than that of $\mathcal{H}_{pl}$:

$$\text{VCD}(\mathcal{H}_{tl}) < \text{VCD}(\mathcal{H}_{pl})$$

The strict inequality holds because $\mathcal{H}_{pl}$ needs the capacity to model potential preference influences from non-target parts, a capacity not required by or included in $\mathcal{H}_{tl}$.

Since the required sample complexity m increases monotonically with the VC dimension for fixed ϵ and δ, and we have established that $\text{VCD}(\mathcal{H}_{tl}) < \text{VCD}(\mathcal{H}_{pl})$, it follows directly that:

$$m_{tl} < m_{pl}$$

Thus, target-focused preference learning is theoretically more sample-efficient than conventional preference learning for achieving the same generalization guarantees regarding the target phenomena.

□

C. Details about dataset construction

This section describes the processes used to construct the dataset and the prompts employed during these processes. The dataset construction consists of five main steps:

- Step 1. Extract images, question-answer pairs associated with the images, and the bounding boxes of objects mentioned in the questions from the Visual Genome dataset.
- Step 2. Use a baseline model to generate responses to the queries.
- Step 3. Compare the model's responses with the answers and filter out only the hallucinated responses that provide incorrect answers.
- Step 4. Compare the images, questions, answers, and hallucinated responses to correct the hallucinated responses into accurate answers.
- Step 5. Compare the hallucinated responses with the revised responses, and retain the revised positions in both the hallucinated responses and revised responses as target positions.

Finally, the dataset we constructed includes images, questions, hallucinated responses, revised responses, target positions, and bounding boxes. In Steps 3 and 4, the ChatGPT model was employed to perform the following tasks:

- Prompt 1. Identifying correctness and errors in the model's responses (in Step 3)
- Prompt 2. Correcting the incorrect model responses (in Step 4)

The prompts used in each step are listed in Table4, providing a comprehensive view of the data generation framework. Examples are also included in the Figure 6.

D. Algorithms

The following (9) is the pseudocode for TL-DPO.

E. Qualitive Results

The following are the results of the qualitative analysis from the hallucination benchmark dataset in figure 7.

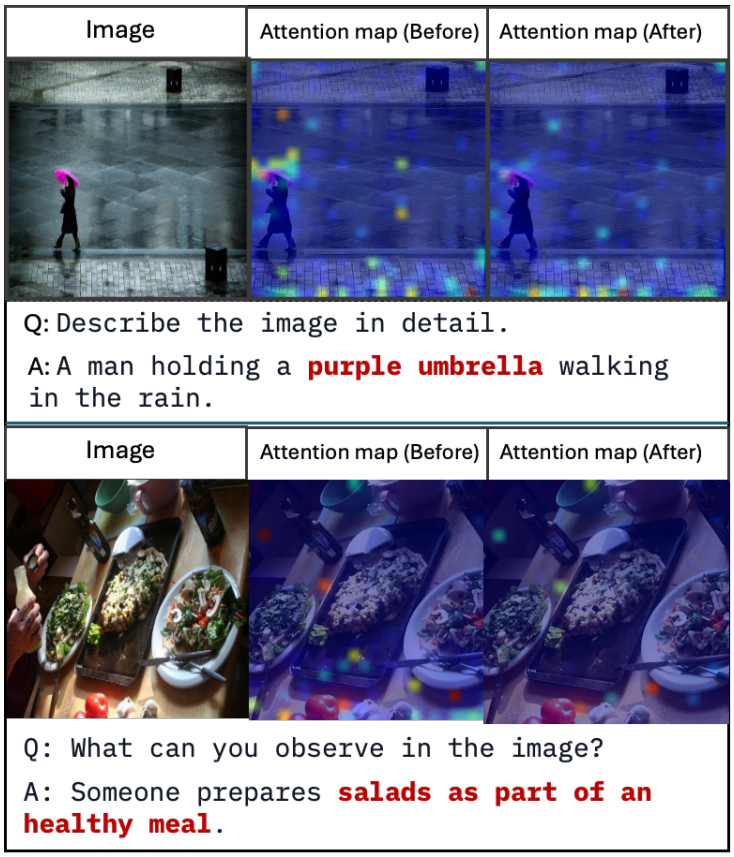

Figure 5. Changes in the Attention Map According to Preference Learning

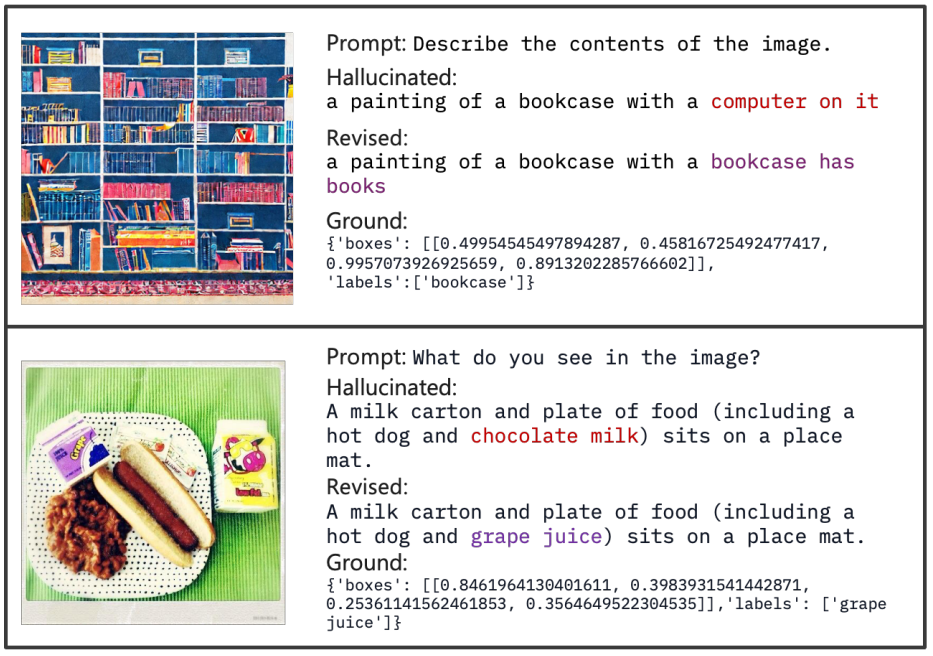

Figure 6. Examples of constructed datasets used in TL-DPO

Algorithm 1 TL-DPO

Require:

- $\mathcal{D}$: Dataset composed of images m, text contexts x, target positions t, and bounding box b.
- π_θ: Parameters of the multimodal language model (MLLM).
- π_{ref}: Parameters of the reference model.
- α, β_1, β_2: Hyperparameters.

Function Definitions:

- LABELTOTEXT(x, t):
  - Labels non-target parts of the text context x to ignore.
  - Truncates the labels after the target position t.
  - Returns the modified labels of text context $l_t(x)$.
- TARGETNOISYMASKING(m, b):
  - Applies a noisy mask to the image m at the regions specified by bounding box b.
  - Returns the modified image $\tilde{m}_t^r$.

Generate TL-DPO data and update $\mathcal{D}$ accordingly.
Initialize model parameters π_θ.
for each epoch **do**
    for each $(m, x, t, b) \in \mathcal{D}$ **do**
        $l_t(x) \leftarrow$ LABELTOTEXT(m, x, k, t) ▷ *Labeling and truncation for $\mathcal{L}_{TL-DPO}$ calculation*
        $\tilde{m}_t \leftarrow$ TARGETNOISYMASKING(m, b) ▷ *Apply noisy masking to target object in image*
        Compute loss $\mathcal{L}_{TL-DPO}$ using Equation (4.3)
        Update π_θ by minimizing $\mathcal{L}_{TL-DPO}$

User: Is the mouse connected to a computer?
LLaVA: Yes, the mouse is connected to a computer, as it is plugged into the computer.
TLDPO: Not connected to any computer.

USER: Which color are the glasses that the man in the image is wearing?
LLaVA: The man in the image is wearing black glasses.'
TLDPO: I can't see the image you're referring to.

USER: How many bicycles are there in the image?
LLaVA: There are two bicycles in the image.
TLDPO: There are three bicycles in the image.

USER: Describe the image in detail.
LLaVA: The image features a group of zebras standing together in a dirt area. There are five zebras in total, with some of them'
TLDPO: The image portrays an interior scene with a focus on a man

USER: Which wooden stool has a vase with red flower on it?
LLaVA: The wooden stool with a vase containing a red flower is on the left side of the image.
TLDPO: The vase with red flowers is on the small wooden table.

Figure 7. Qualitative Evaluation of Hallucination Correction Performance

Table 4. Two types of prompts to GPT-4o (used in Step3, Step4)

Identifying correctness in the model's responses:

Help me evaluate the correctness of the model's responses by comparing them to the dataset's Question-Answer pair.

****

Question-answer Pair:

Q: {question}

A: {answer}

Response:

R: {response}

Requirements:

(1) Compare the Response to the Question with the Answer and determine its correctness.

(2) Provide the result as either "true" or "false".

****

Output Format:

Output: {your answer}

Prompts for revising hallucination tasks:

Requirements:

(1) Modify parts of the given incorrect response to make it a correct response.

(2) Compared to the original output, the modified response should be corrected based on the provided image and the correct answer.

(3) Highlight the corrected parts by wrapping them with asterisks (e.g., corrected text).

********** Output Format:

Revised answer: {your answer}