

# Supplementary Materials: AIGCs Confuse AI Too: Investigating and Explaining Synthetic Image-induced Hallucinations in Large Vision-Language Models

Anonymous Authors

## 1 SEMANTICS TRANSLATION

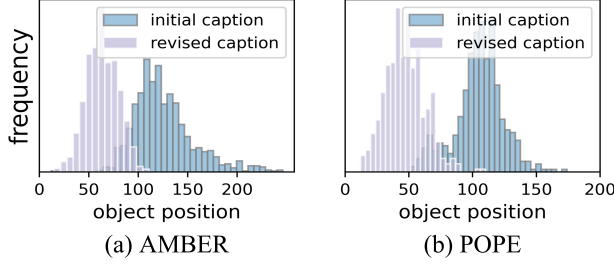


Figure 1: The comparison of object positions before and after the caption revision. Taking Stable Diffusion as an example, where the accepted character limit is 77, the distribution of key semantics in the revised caption generally satisfies the character limits.

### 1.1 Instruction of Caption Revision

This section provides a detailed instruction for caption revision. Since ChatGPT cannot perceive visual signals, we provide it with the manual annotation results to aid its understanding of the draft caption. Specifically, we expect the model to extract key elements, including scenes, existent objects, their attributes, and other crucial semantic information. Moreover, to mitigate the impact of redundant or speculative descriptions in the draft captions on image generation, we aim for the corrected captions to omit unimportant information like emotional expressions or irrelevant associations. Ultimately, our goal is to generate a concise and easily understandable English image caption using simple vocabulary, not exceeding 80 words. As shown in Figure 3, the revised caption successfully captures the attribute semantics of objects in the image, such as *'three people walking.'* Compared to the original *'some individuals walking,'* the former better aligns with the visual semantics and can accurately prompt the generation model. Furthermore, the revised caption removes redundant information, such as *'indicating that they might be enjoying a recreational activity,'* which provides no additional benefit to image synthesis. We also compare the distribution of key objects in caption before and after revision, as shown in Figure 1, the length of the revised caption is generally in line with the word limit set by the generative model, ensuring that all key semantic information can effectively prompt the generative model.

### 1.2 Segmentation Tools in Image Filtering

To ensure authentic semantics in synthetic images, we mainly focus on avoiding (i) the depiction of objects not existing in natural images or (ii) introducing objects that contradict human cognition. Thus we initially extract objects using automated segmentation tools [5]

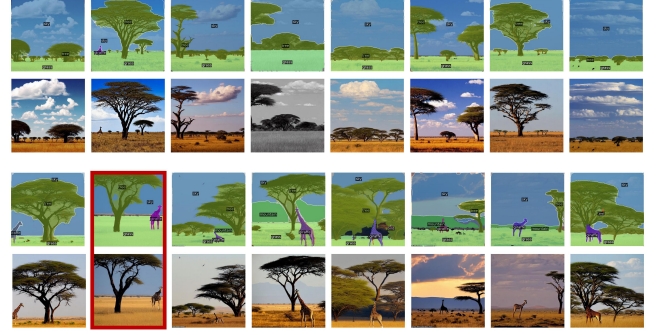


Figure 2: The segmentation results on the candidate set of synthetic images. The annotation results of corresponding natural image is: *"sky, tree, giraffe and grass"*. The red box indicates the final selected synthetic image.

(SEEM) and compare the consistency of the extracted objects on natural and synthetic images. SEEM is a promptable and interactive model for segmenting everything everywhere all at once in an image. Specifically, we highlight its capability in open-vocabulary segmentation. Figure 2 demonstrates a case of segmentation results on the candidate set of synthetic images. We then eliminate images displaying an excess or absence of objects in their annotation results when compared to the corresponding natural image.

### 1.3 Similarity Calculation in Image Filtering

**Metrics on Perceptual Similarity:** We use DreamSim[2] to measure the perceptual similarity between the synthetic and its corresponding natural image. We denote a distance between two images as  $D(\cdot, \cdot; f_\theta)$ , where  $f_\theta$  is a feature extractor. We consider the ensemble of three transformer-based backbones: DINO[1], CLIP[4] and MAE[3]. Following the setting in DreamSim, the distance  $D(x, \hat{x}; f_\theta) = 1 - \cos(f_\theta(x), f_\theta(\hat{x}))$  is taken as the cosine distance between the  $[CLS]$  tokens taken from the last layer for DINO and MAE, and the embedding vector for CLIP.

**Metrics on Semantic Faithfulness:** With the help of manual annotations, we construct a batch of text descriptions for existing objects. Specifically, we employ the common prompt template in CLIP: "There is a photo of {object}". Subsequently, we utilize the CLIP model to obtain the vision embedding  $E_v$  and text embeddings  $\{E_{t_1}, E_{t_2}, \dots, E_{t_n}\}$ , where  $n$  denotes the number of text descriptions. We then calculate the average cosine similarity between each text description and the given synthetic image using the CLIP model:  $S_{CLIP} = 1/n \sum_{i=1}^n \cos(E_v, E_{t_i})$ .

## REFERENCES

- [1] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. 2021. Emerging properties in self-supervised

**system message:** 你是一个优秀的图像字幕（Image caption）助理，图像字幕标注是计算机对视觉图像感知能力的体现，要求客观准确的反映图像中存在的物体，并通过自然语言的方式表达出来。

**Instruction:** 你将收到一段真实图像的英文描述，注意此时并不是合格的图像字幕，但包含了组成图像字幕的关键信息；你还将收到真实图像中，人工标注的物体信息集合Sobj，你的要求如下：

- 1、你需要借助Sobj中的准确信息辅助你理解这段英文描述，并将其转化为图像字幕；
- 2、你需要提取其中的关键信息，包括场景、出现的物体等重要信息；
- 3、合成字幕时请缩减不重要的信息，例如带有推测或者情感表达的语句；
- 4、你需要精简你的输出结果，用尽可能简单易懂的词汇合成一段不超过80词的英文图像字幕。

**## Input**

**Description:** "The image shows some individuals walking through a grassy field on a sunny day. The sky is deep blue and clear, with no visible clouds. The landscape in the distance includes rolling hills or mountains, suggesting that the field is located at a higher elevation or in a hilly region. The greenery is lush and the day appears to be quite pleasant for an outdoor walk or hike. The people are dressed in casual outdoor clothing, indicating that they might be enjoying a recreational activity. The field is bordered by taller grasses or shrubs and there is a visible path where the individuals are walking. The scenery suggests it could be a rural or natural park setting."

**Sobj:** {sky, forest, grass, person, lake, mountain, road, the grass is green in this image, there is a direct contact between the person and grass, there is no direct contact between the person and tree.}

**## Output**

**Image Caption:** "The image captures three people walking in a lush, green field on a sunny day. The deep blue, cloudless sky and distant rolling hills or mountains suggest a higher elevation. The individuals, dressed casually, appear to be enjoying a pleasant outdoor walk. The field, bordered by tall grasses, features a visible path, creating a rural or natural park setting."

Figure 3: Instructions used to prompt ChatGPT-3.5 for caption revision. We utilize Chinese to describe the instructions and require ChatGPT to revise the given draft caption based on the manual annotation results.

vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*. 9650–9660.

[2] Stephanie Fu, Netanel Tamir, Shobhita Sundaram, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. 2023. DreamSim: Learning New Dimensions of Human Visual Similarity using Synthetic Data. *arXiv preprint arXiv:2306.09344* (2023).

[3] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. 2022. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 16000–16009.

[4] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PMLR, 8748–8763.

[5] Xueyan Zou, Jianwei Yang, Hao Zhang, Feng Li, Linjie Li, Jianfeng Wang, Lijuan Wang, Jianfeng Gao, and Yong Jae Lee. 2024. Segment everything everywhere all at once. *Advances in Neural Information Processing Systems* 36 (2024).