When Lighting Deceives: Exposing Vision-Language Models' Illumination Vulnerability Through Illumination Transformation Attack

Supplementary Material

Overview

This supplementary material provides essential details that complement our main paper. Sec. A presents the ITA update formulation, which builds upon the canonical form of CMA-ES [1, 2]. Sec. B lists the specific class names of the 30 selected categories in the COCO dataset. Sec. C provides all prompt templates given to GPT-4o and GPT-4 for the Illumination Natural Score and consistency and correctness metrics. Sec. E showcases additional visualizations of illumination-aware adversarial examples and their performance across various VLMs.

A. Details of Optimization Algorithm

Our method optimizes adversarial illumination distributions through the Covariance Matrix Adaptation Evolution Strategy (CMA-ES) [1, 2]. Among various evolutionary optimization algorithms, CMA-ES stands out as one of the most effective approaches, demonstrating superior performance particularly on medium-scale optimization problems (typically involving 3-300 variables) [1]. Its gradient-free nature eliminates the dependency on gradient information, making it an ideal choice for optimizing the adversarial illumination distributions in our framework. Additionally, we employ Learning Rate Adaptation (LRA) and Early Stopping policy for efficient search. These enhancements improve convergence speed and prevent unnecessary iterations, making the approach suitable for real-world applications. The adversarial illumination configuration $\mathbf{\Lambda}$ is parameterized as follows:

$$\mathbf{\Lambda} = \mathbf{A} \cdot \tanh(\mathbf{q}) + \mathbf{B}, \quad \text{where} \quad \mathbf{q} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{C}\mathbf{\Sigma}^{\mathbf{2}}), \tag{A.1}$$

where $\mathbf{A}$ and $\mathbf{B}$ scale Gaussian samples into the feasible range, and $\mathbf{q}$ follows a Gaussian distribution with mean $\boldsymbol{\mu} \in \mathbb{R}^{4n}$, step-size $\mathbf{\Sigma} \in \mathbb{R}_{>0}$, and covariance matrix $\boldsymbol{C} \in \mathbb{R}^{4n \times 4n}$. The optimization maximizes the adversarial impact on VLMs:

$$\arg \max_{\boldsymbol{\mu}, \boldsymbol{C}, \mathbf{\Sigma}} \mathbb{E}_{\mathbf{q} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{C}\mathbf{\Sigma}^{\mathbf{2}})} \left[\mathcal{L}_{\text{Adv}}(X', Y) + \alpha \cdot \mathcal{L}_{\text{Pecp}} + \beta \cdot \mathcal{L}_{\text{Dis}} \right], \quad \text{where} \quad X' = \mathcal{I}(X, \mathbf{A} \cdot \tanh(\mathbf{q}) + \mathbf{B}). \tag{A.2}$$

The optimization follows three main steps:

1). Sampling: Generate a population of candidate adversarial illumination parameters from the Gaussian distribution:

$$\mathbf{q}^{(i)} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{C}\mathbf{\Sigma}^{\mathbf{2}}), \quad i = 1, \ldots, K. \tag{A.3}$$

These samples are then transformed into valid illumination configurations:

$$\mathbf{\Lambda}^{(i)} = \mathbf{A} \cdot \tanh(\mathbf{q}^{(i)}) + \mathbf{B}. \tag{A.4}$$

2). Evaluation: Compute the objective function values for each sample:

$$f^{(i)} = \mathcal{L}_{\text{Adv}}(X'^{(i)}, Y) + \alpha \cdot \mathcal{L}_{\text{Pecp}} + \beta \cdot \mathcal{L}_{\text{Dis}}, \tag{A.5}$$

where $X'^{(i)} = \mathcal{I}(X, \mathbf{\Lambda}^{(i)})$ represents the relit image under the adversarial illumination conditions, and α and β denote the weights.

3). Update: The distribution parameters $\boldsymbol{\mu}$, $\boldsymbol{C}$, and $\mathbf{\Sigma}$ are updated using the CMA-ES strategy. In addition, the learning rate adaptation (LRA) and early stopping policy are applied during the update process to enhance convergence and prevent unnecessary iterations.

$$\boldsymbol{\mu} \leftarrow \sum_{i=1}^{K} w_i \mathbf{q}^{(i)}, \tag{A.6}$$

$$\boldsymbol{C} \leftarrow (1 - c_c)\boldsymbol{C} + c_c \sum_{i=1}^{K} w_i (\mathbf{q}^{(i)} - \boldsymbol{\mu})(\mathbf{q}^{(i)} - \boldsymbol{\mu})^T, \tag{A.7}$$

$$\mathbf{\Sigma} \leftarrow \mathbf{\Sigma} \cdot \exp\left(c_\sigma \left(\frac{\|\mathbf{p}\|}{\mathbb{E}[\|\mathcal{N}(0, I)\|]} - 1 \right) \right), \tag{A.8}$$

where w_i are the selection weights, c_c and c_σ are learning rates, and $\mathbf{p}$ is the evolution path for step-size control. And the **Learning Rate Adaptation (LRA)** is integrated into the update of $\mathbf{\Sigma}$:

$$\mathbf{\Sigma}_{\text{new}} \leftarrow \mathbf{\Sigma} \cdot \exp\left(\frac{c_\sigma}{\|\mathbf{p}\|} \right), \tag{A.9}$$

where $\|\mathbf{p}\|$ is the norm of the evolution path and c_σ is a learning rate controlling the adaptation. Additionally, the **Early Stopping Policy** is implemented by monitoring the following conditions during the optimization process:

$$\Delta f_{\text{best}} = |f_{\text{best}}^{(i)} - f_{\text{best}}^{(i-1)}| < \delta, \tag{A.10}$$

where $f_{\text{best}}^{(i)}$ is the best fitness at iteration i and δ is a small threshold. If the fitness change is below this threshold, early

Table S.1. The selected 30 categories in COCO dataset.

0	airplane	1	banana	2	bear
3	bed	4	bird	5	boat
6	broccoli	7	bus	8	cake
9	cell phone	10	clock	11	cow
12	dog	13	donut	14	elephant
15	fire hydrant	16	horse	17	kite
18	motorcycle	19	pizza	20	sandwich
21	teddy bear	22	traffic light	23	stop sign
24	toilet	25	train	26	umbrella
27	vase	28	zebra	29	sheep

stopping is triggered. The optimization will stop if the number of iterations exceeds a predefined maximum limit or if there is no improvement in the best fitness value for a specified number of consecutive generations.

B. Selected COCO Categories

We conduct experiments of zero-shot classification task (Tab. 1) on 30 COCO Categories, generating Adv-IT samples from COCO validation set images. The selected categories are enumerated in Tab. S.1.

C. Prompt Templates

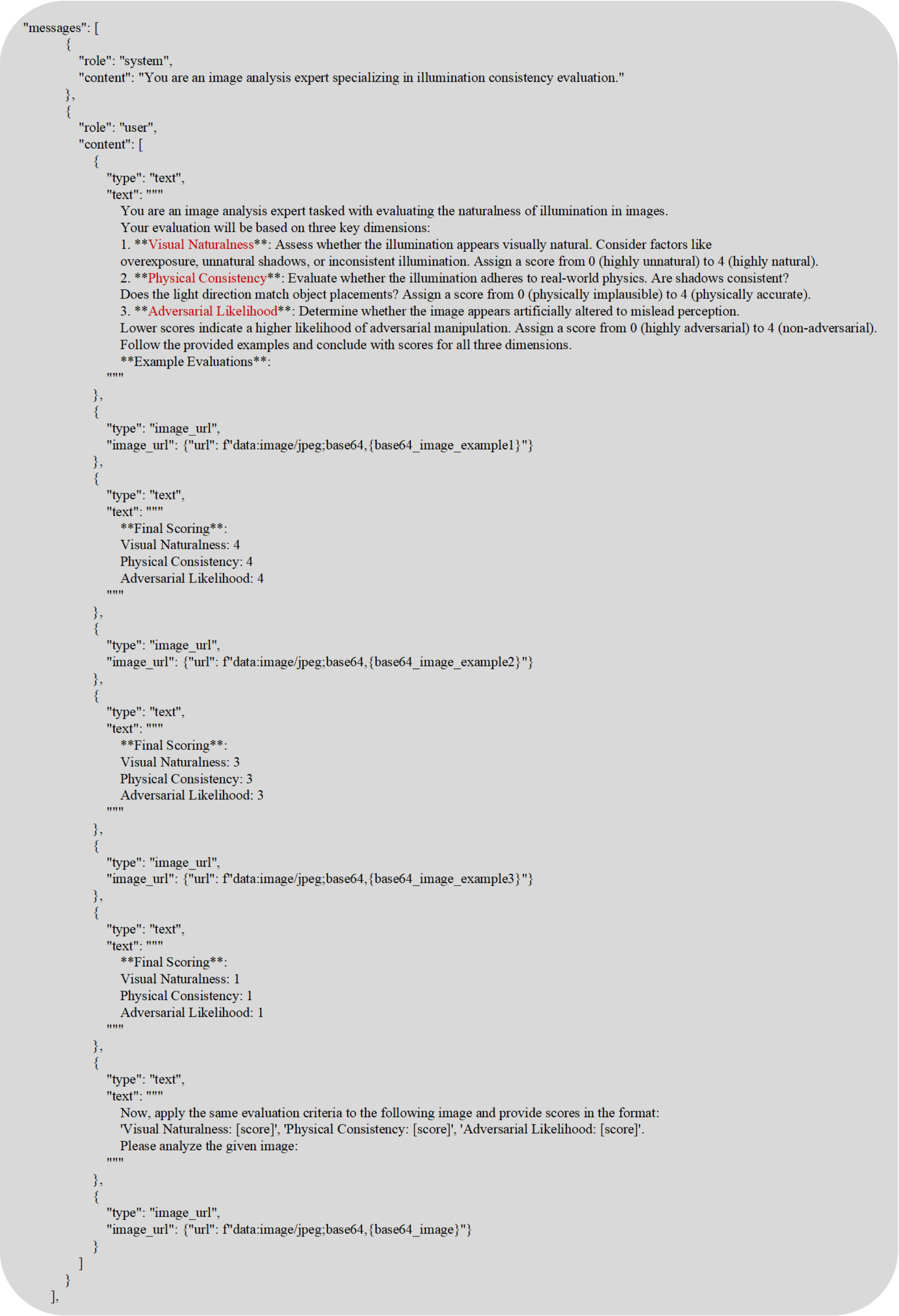
Fig. S.1 illustrates the prompt template used for evaluating image illumination naturalness in our main experiments (Tab. 1). The prompt templates for computing GPT-Score, which are employed to assess image captioning and VQA performance in our main experiments (Tab. 2 and Tab. 3), are presented in Fig. S.2 and Fig. S.3. Notably, to unify the metrics, these two scores were converted to their respective percentages.

D. Computational Cost

The computational cost of optimization for each clean sample is approximately 40 GPU minutes. Our experiments on the COCO validation set (300 samples) took 24 GPU hours on 8 NVIDIA RTX 4090 GPUs.

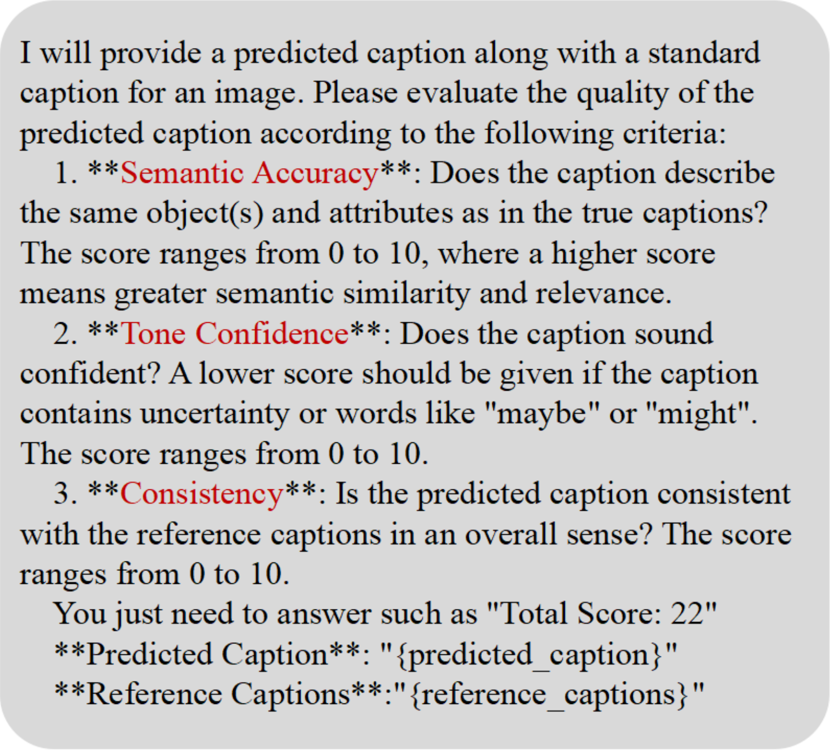
E. More Visualization Examples

We provide additional visualization examples of illumination-aware adversarial examples generated by ITA in Fig. S.4.



```
"messages": [
        {
          "role": "system",
          "content": "You are an image analysis expert specializing in illumination consistency evaluation."
        },
        {
          "role": "user",
          "content": [
            {
              "type": "text",
              "text": """
                You are an image analysis expert tasked with evaluating the naturalness of illumination in images.
                Your evaluation will be based on three key dimensions:
                1. **Visual Naturalness**: Assess whether the illumination appears visually natural. Consider factors like
                overexposure, unnatural shadows, or inconsistent illumination. Assign a score from 0 (highly unnatural) to 4 (highly natural).
                2. **Physical Consistency**: Evaluate whether the illumination adheres to real-world physics. Are shadows consistent?
                Does the light direction match object placements? Assign a score from 0 (physically implausible) to 4 (physically accurate).
                3. **Adversarial Likelihood**: Determine whether the image appears artificially altered to mislead perception.
                Lower scores indicate a higher likelihood of adversarial manipulation. Assign a score from 0 (highly adversarial) to 4 (non-adversarial).
                Follow the provided examples and conclude with scores for all three dimensions.
                **Example Evaluations**:
              """
            },
            {
              "type": "image_url",
              "image_url": {"url": f"data:image/jpeg;base64,{base64_image_example1}"}
            },
            {
              "type": "text",
              "text": """
                **Final Scoring**:
                Visual Naturalness: 4
                Physical Consistency: 4
                Adversarial Likelihood: 4
              """
            },
            {
              "type": "image_url",
              "image_url": {"url": f"data:image/jpeg;base64,{base64_image_example2}"}
            },
            {
              "type": "text",
              "text": """
                **Final Scoring**:
                Visual Naturalness: 3
                Physical Consistency: 3
                Adversarial Likelihood: 3
              """
            },
            {
              "type": "image_url",
              "image_url": {"url": f"data:image/jpeg;base64,{base64_image_example3}"}
            },
            {
              "type": "text",
              "text": """
                **Final Scoring**:
                Visual Naturalness: 1
                Physical Consistency: 1
                Adversarial Likelihood: 1
              """
            },
            {
              "type": "text",
              "text": """
                Now, apply the same evaluation criteria to the following image and provide scores in the format:
                'Visual Naturalness: [score]', 'Physical Consistency: [score]', 'Adversarial Likelihood: [score]'.
                Please analyze the given image:
              """
            },
            {
              "type": "image_url",
              "image_url": {"url": f"data:image/jpeg;base64,{base64_image}"}
            }
          ]
        }
      ],
```

Figure S.1. The prompt template for image illumination naturalness evaluation by GPT-4o



I will provide a predicted caption along with a standard caption for an image. Please evaluate the quality of the predicted caption according to the following criteria:

1. **Semantic Accuracy**: Does the caption describe the same object(s) and attributes as in the true captions? The score ranges from 0 to 10, where a higher score means greater semantic similarity and relevance.

2. **Tone Confidence**: Does the caption sound confident? A lower score should be given if the caption contains uncertainty or words like "maybe" or "might". The score ranges from 0 to 10.

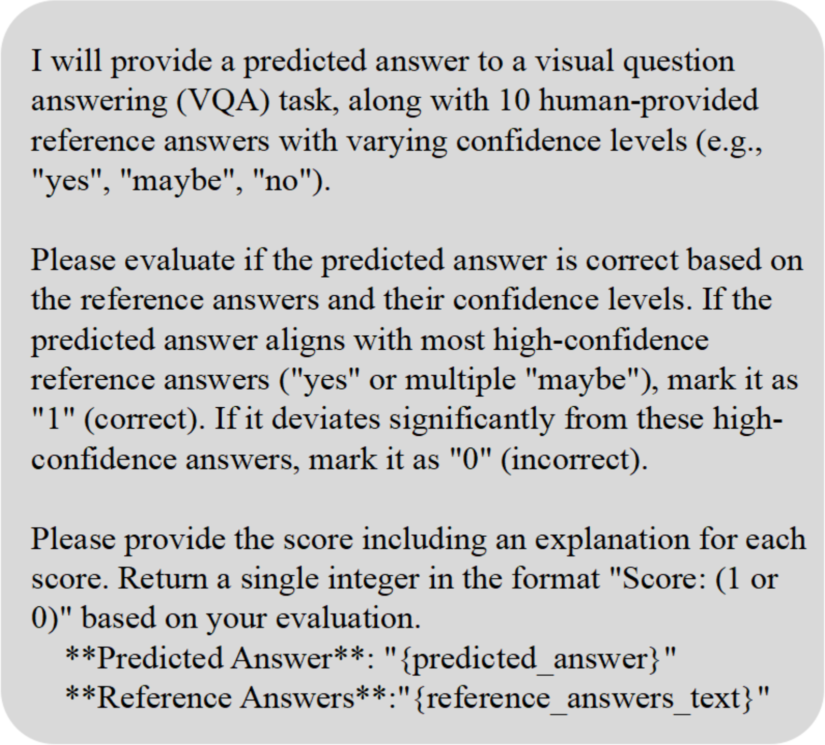
3. **Consistency**: Is the predicted caption consistent with the reference captions in an overall sense? The score ranges from 0 to 10.

You just need to answer such as "Total Score: 22"

Predicted Caption: "{predicted_caption}"

Reference Captions:"{reference_captions}"

Figure S.2. The prompt template for consistency in image captioning tasks.



I will provide a predicted answer to a visual question answering (VQA) task, along with 10 human-provided reference answers with varying confidence levels (e.g., "yes", "maybe", "no").

Please evaluate if the predicted answer is correct based on the reference answers and their confidence levels. If the predicted answer aligns with most high-confidence reference answers ("yes" or multiple "maybe"), mark it as "1" (correct). If it deviates significantly from these high-confidence answers, mark it as "0" (incorrect).

Please provide the score including an explanation for each score. Return a single integer in the format "Score: (1 or 0)" based on your evaluation.

Predicted Answer: "{predicted_answer}"

Reference Answers:"{reference_answers_text}"

Figure S.3. The prompt template for correctness in VQA tasks.

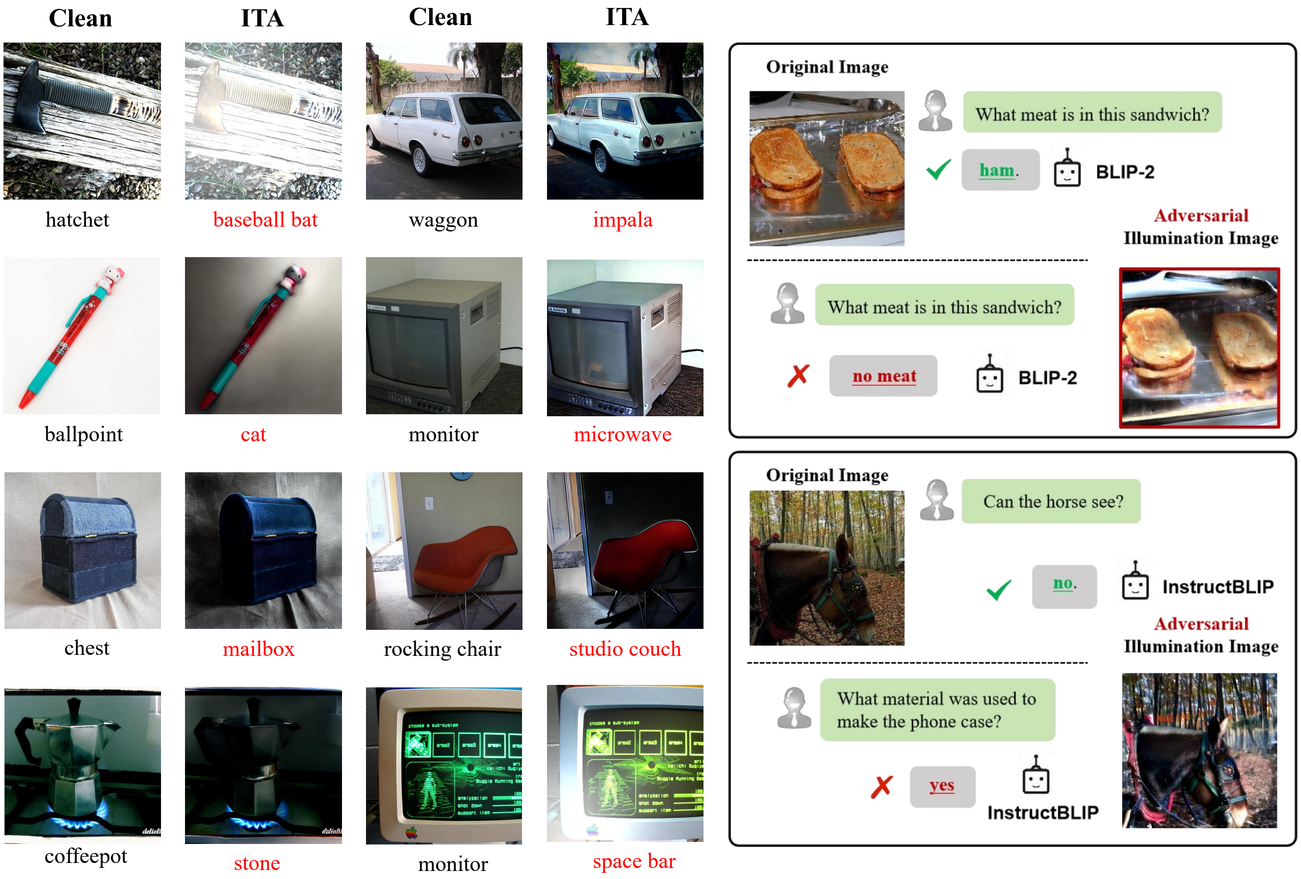


Figure S.4. Additional visualization of illumination-aware adversarial examples.

References

[1] Daniel Golovin, Benjamin Solnik, Subhodeep Moitra, Greg Kochanski, John Karro, and David Sculley. Google vizier: A service for black-box optimization. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1487–1495, 2017. 1

[2] Nikolaus Hansen. The cma evolution strategy: A tutorial. *arXiv preprint arXiv:1604.00772*, 2016. 1