DualMirage: Hunting Stealthy Multimodal LLM Agents via CAPTCHAs with Contour and Adversarial Illusions

Supplementary Material

8. Implementation Details

Contour-illusory image configurations. We demonstrate the effectiveness and generality of our proposed Colored Abutting Grating Illusion Generation algorithm through two distinct image types: anime-style portraits (Illusion-Selfie2Anime) and handwritten digits (Illusion-MNIST).

Figure 5 showcases contour-illusory images generated from the Selfie2Anime dataset. Real portraits and anime-style portraits from the Selfie2Anime dataset were processed at two different resolutions for comparative analysis. We experimented with grating distortions at two image resolutions: 64×64 and 256×256. Our preliminary experiments revealed that MLLMs exhibit significantly poorer perception of contour illusions in low-resolution distorted images (64×64). Therefore, we employed 64×64 images for the contour illusion component in our main experiments. However, for the subsequent adversarial example generation phase, we resize the images to 256×256 to expand the search space and enhance the effectiveness of adversarial perturbations. We systematically evaluated different grating orientations and found that vertical gratings ($\theta = 90°$) most effectively disrupt facial features in cartoon character images, as they create strong interference patterns that break the continuity of horizontal facial contours. In experimentation, we observed that larger grating periods cause more severe distortion to the original image content. However, excessively large periods also degrade human readability. After balancing these competing factors, we selected $T = 4$ pixels as the optimal compromise, providing sufficient distortion to challenge MLLMs while maintaining high human recognition accuracy.

Figure 6 demonstrates examples of contour-illusory images generated from the MNIST handwritten digit dataset. MNIST grayscale digit images with varying sequence lengths $l \in \{3, 6, 10\}$. Set to $T = 4$ pixels for optimal illusion strength, as our preliminary experiments found that this spatial frequency creates strong contour illusions while maintaining human readability. For the digit recognition task, we simplified the hyperparameter configuration by fixing the grating orientation to horizontal ($\theta = 0°$) and directly using the original MNIST dataset resolution (28×28 pixels) without resizing. Fixed at horizontal orientation ($\theta = 0°$) to maintain consistency across all digit recognition experiments. Although the adversarial example generation showed limited effectiveness in this configuration, the contour illusion component demonstrated remarkable capability in confusing MLLMs. Our experiments revealed that MLLMs often pretend to recognize digits and fabricate random numbers as responses. This distinctive behavioral pattern, where models confidently generate incorrect digit sequences rather than admitting inability to perceive the contours, provides a strong discriminative signal for distinguishing humans from MLLM agents. The high agent blocking rates (up to 100% across all tested models) confirm the effectiveness of this approach despite the simplified parameter setup.

Computational cost. The PGD-based attacks against various backbones, including ViT-B/32, ViT-B/16, and ViT-L/14, are conducted offline to sufficiently explore the adversarial space. This process typically requires 1–5 hours of runtime using two NVIDIA A100 GPUs. As our method is designed for proactive image protection prior to distribution, it does not require real-time generation during inference.

9. Further Empirical Analysis

In this section, we provide additional experimental results and in-depth analyses to further validate the robustness and generalizability of our proposed DualMirage.

Differential sensitivities of closed-source models. As discussed in Sec. 5.2, commercial MLLMs exhibit varying degrees of sensitivity to contour illusions. Specifically, GPT models demonstrate the highest robustness against contour interference, often maintaining the ability to recognize figures at higher resolutions. Gemini's recognition performance shows a positive correlation with increased image brightness, whereas Claude appears to be the most vulnerable to contour-based disruptions. Interestingly, despite these variations in handling contour illusions, all evaluated closed-source models exhibit highly comparable vulnerabilities to our generated adversarial illusions. Visual examples highlighting these distinct failure modes are provided in Figure 4.

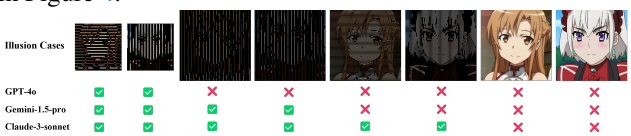

Figure 4. **Visual examples of differential model sensitivities.** This figure illustrates how different closed-source MLLMs respond distinctively to contour illusions while showing comparable vulnerability to adversarial illusions.

Generalizability on complex visual recognition tasks. The primary focus of this work is the novel and previously

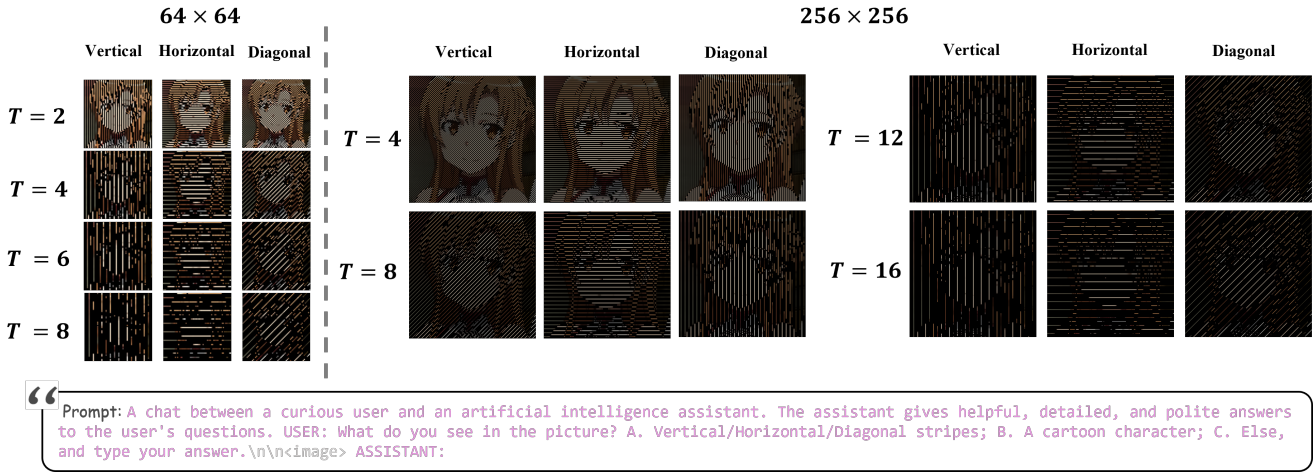

Figure 5. **Contour-illusory image examples from Illusion-Selfie2Anime challenges.** Original anime portraits are processed into contour-illusory versions using varying resolutions (64×64, 256×256), grating orientations ($\theta \in \{0°, 45°, 90°\}$), and periods ($T$).

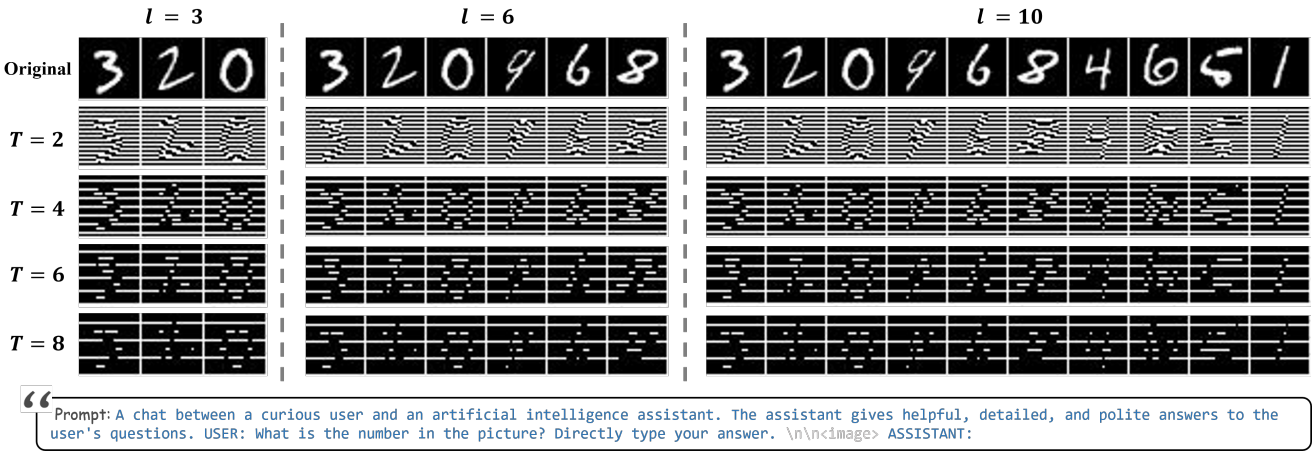

Figure 6. **Contour-illusory image examples from Illusion-MNIST challenges.** The original images (row 1) show clear grayscale digit images, while the processed outputs (row 2-5) display contour-illusory images composed of black and white stripes. Humans can easily identify digit sequences (e.g., "3 2 0 9 6 8") in the right column, while MLLM models struggle to accurately perceive these illusory-formed contours.

unexplored task of "hunting stealthy agents." Due to the lack of existing benchmarks for this specific objective, we introduced two pioneering datasets, Illusion-Selfie2Anime and Illusion-MNIST, tailored for this task. To further validate the broad applicability and generalizability of our proposed DualMirage on more complex and realistic visual understanding challenges, we extended our evaluation to the Illusion-ImageNet dataset (derived from ImageNet-100). This dataset presents a much richer semantic space compared to digit or portrait datasets. As demonstrated in Table 5, DualMirage maintains a high level of effectiveness on Illusion-ImageNet. The strong performance across human usability (*HSR*), agent blocking capability (*ABR*), and adversarial attack capability (*ASR*) confirms that our approach successfully scales beyond structurally simple datasets and remains robust in semantically complex, real-world visual scenarios.

Table 5. **Performance of DualMirage on the Illusion-ImageNet challenge.** The results demonstrate the effectiveness of our method on tasks with high semantic complexity.

Task	HSR	Avg. ABR (%)	Avg. ASR (%)
Illusion-ImageNet	98.2	93.3	45.7