

# Supplementary for Web Artifact Attacks Disrupt Vision Language Models

Maan Qraitem, Piotr Teterwak, Kate Saenko, Bryan A. Plummer  
Boston University  
{mqraitem, piotrt, saenko, bplum}@bu.edu

## A. More examples of Attack Artifacts

In Sec. 3.2, we discuss insights from the found attack artifacts. In this section, we offer more examples and discussion. Fig. 1 reveals consistent patterns in the types of artifacts that successfully manipulate model predictions. Text-based artifacts frequently include words that share phonetic, visual, or partial textual similarities with the target class. For instance, “KID” appears as an artifact for Child, likely due to its strong semantic association, while “BING” is linked to Boeing, exploiting visual resemblance to the company name. Similarly, artifacts like “K-POP” for South Korea and “Small” for Smiling demonstrate how models latch onto common co-occurring words rather than genuine visual cues.

Graphics with embedded text also play a significant role in misleading models. These artifacts often feature brand names, stylized logos, or generic text associated with the target category. For instance, the inclusion of “Abacus” and “Arbys” as artifacts for Airbus suggests that the model has learned to associate certain brand names or typographic styles with aircraft manufacturers. Likewise, “Happy Bakery Cafe” and “Smile!” appearing under Smiling indicate that models are influenced by commercial logos and positive branding elements rather than facial features.

Graphics without text rely purely on visual resemblance, symbols, and branding elements that loosely connect to the target category. For example, aviation-related symbols such as the Airbus logo and Wi-Fi icon appear under Airbus, and basketball team logos (*e.g.*, Portland Trail Blazers, Atlanta Hawks) emerge under Boeing, possibly due to their circular and wing-like shapes. Similarly, for Man and Woman, symbols traditionally linked to gender (*e.g.*, mustaches, gendered icons, and heart-shaped logos) suggest that models encode stereotypical visual representations rather than deeper semantic understanding.

Across all target classes, these results highlight that Web Artifact Attacks exploit both direct textual similarities and broader visual associations, making them an effective and adaptable attack vector. The presence of corporate logos, branding, and culturally specific symbols (*e.g.*, the Brazilian flag colors, K-Pop branding for South Korea) suggests that models are influenced by common internet-scale data dis-

tributions rather than purely semantic understanding. This demonstrates the pervasive reliance on spurious correlations, emphasizing the need for more robust dataset curation and training strategies to mitigate these vulnerabilities.

## B. Attack Success Rate by Location

In Sec. 3.1.3, we discussed how we optimize the artifact location placement as part of our attack. Fig. 2 examines how the placement of artifacts affects attack success rates across different artifact types. The results indicate that artifacts positioned in the top-center region of the image consistently lead to higher misclassification rates, with text-based artifacts being the most effective. This suggests that vision-language models prioritize information in certain spatial regions, likely due to biases in pretraining datasets, where text frequently appears near the top of images (*e.g.*, headlines, labels, or watermarks). Graphics with embedded text also show higher success rates in the top-center, though to a lesser extent than pure text, while graphics without text have a more uniform but overall lower effect. These findings highlight the importance of spatial biases in model vulnerability and suggest that adversarial manipulations may be optimized further by strategically placing artifacts in regions that models inherently attend to more strongly.

## C. Attack Success by Dataset

In Sec. 4.1, we report the average performance over 5 datasets. In this Section, we break down performance by dataset. Fig. 3 shows that Web Artifact Attacks achieve higher success rates in datasets containing human-related attributes (FairFace [2], CelebA [3]) compared to non-human classification tasks (Aircraft [4], Country211 [6]). This trend is particularly pronounced for text-based artifacts, which consistently lead to the highest misclassification rates in human-related datasets. One possible explanation is that vision-language models are pretrained on web-scale data where text often co-occurs with human-related concepts, reinforcing spurious associations between textual artifacts and human characteristics. In contrast, non-human tasks like aircraft recognition rely more on fine-grained visual details,





























































Target Class	$A_{c^*}^{\text{graphics-no-text}}$	$A_{c^*}^{\text{graphics-text}}$	$A_{c^*}^{\text{text}}$
Brazil	  	  	Erulas Barel brave
South Korea	  	  	Kong CODE K-POP
Teen	  	  	trigger Tee Youth
Child	  	  	CHID KID KIdZ
Man	  	  	Mr; Bean Mr Father
Woman	  	  	Woinen VODKA Kivoria
Boeing	  	  	bofing Beang BING
Airbus	  	  	Areus ABUS ATRIIS
Smiling	  	  	Small Spnle SHAPE
Frowning	  	  	FROG Broken BURG

Figure 1. **Examples of Attack Artifacts** categorized into three types: graphics without text ( $A_{c^*}^{\text{graphics-no-text}}$ ), graphics with embedded text ( $A_{c^*}^{\text{graphics-text}}$ ), and unrelated text ( $A_{c^*}^{\text{text}}$ ). Each row corresponds to a different target class showing artifacts that models have learned to associate with these concepts. Notably, text artifacts need not match the class exactly, while graphical symbols can represent indirect but learned associations. These findings highlight the diverse range of artifacts that can manipulate model predictions.

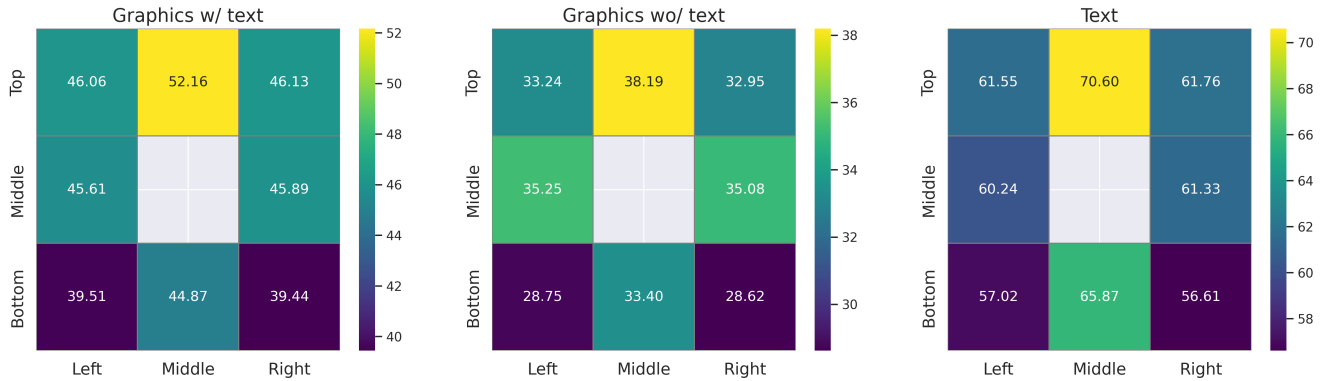


Figure 2. **Effect of artifact placement on attack success rates** Heatmaps show success rates when artifacts are positioned in different regions of the image for three artifact types: graphics with text (left), graphics without text (middle), and text (right). Across all artifact types, placing the artifact in the top-center region consistently yields the highest success rates, particularly for text-based attacks, which reach up to 70% success in this position.



Figure 3. **Artifact attack success rates across datasets with human-related and non-human-related tasks.** Each subplot corresponds to a different artifact type: graphics with text (left), graphics without text (middle), and text (right). Across all artifact types, human-related tasks (e.g., FairFace, CelebA) exhibit higher vulnerability compared to non-human tasks (e.g., Aircraft, Country211), with text-based artifacts being the most effective overall.

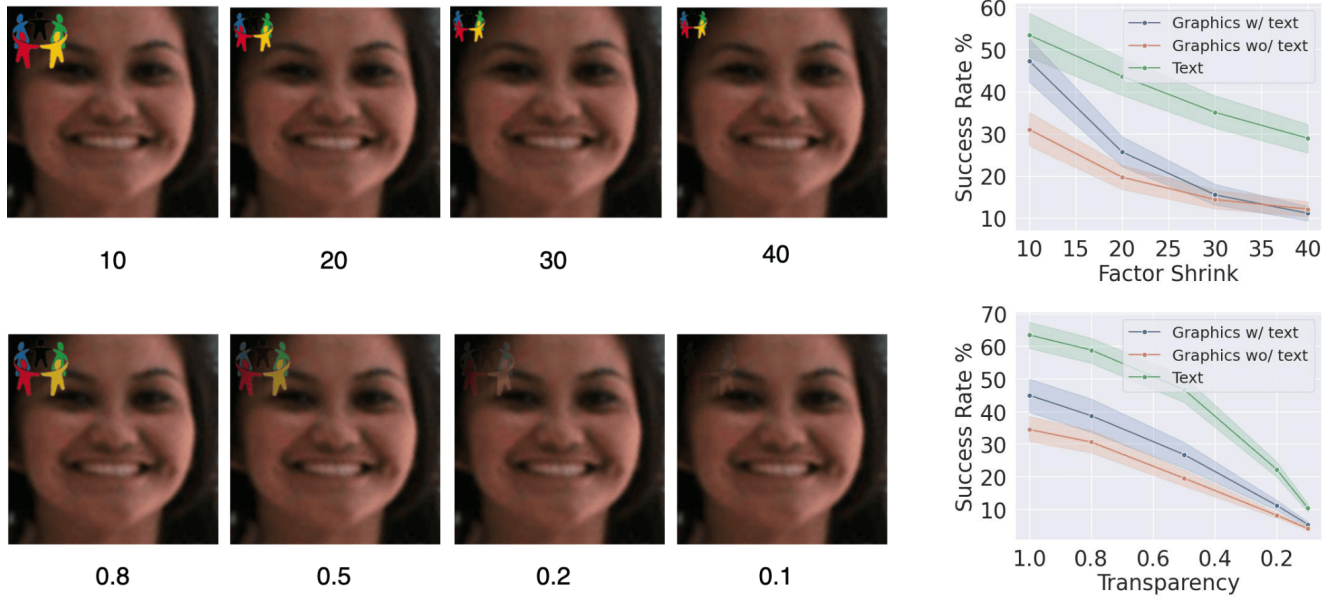


Figure 4. **Effect of artifact size and transparency on attack success rates.** The top row shows examples of artifacts shrinking in size (Factor Shrink) from 10 to 40, while the bottom row illustrates decreasing artifact opacity (Transparency) from 1.0 (fully visible) to 0.1 (highly transparent). The corresponding line plots on the right show the attack success rates across different artifact types. Smaller and more transparent artifacts consistently reduce attack effectiveness, but text-based artifacts remain the most effective even under these constraints.

making them less susceptible to artifacts that exploit text or graphical elements. These findings suggest that Web Artifact Attacks pose a greater risk to applications involving human-centric classifications, where models may rely more heavily on text-based biases.

#### D. Effect of Artifact’s Transparency and Size

In Sec. 4.1., we fix the artifact size to 10th of the image sizes and transparency to 1.0. In this Section, we ablate

both settings. Fig. 4 evaluates how reducing artifact size and transparency affects attack success rates. The results show that as artifacts become smaller or more transparent, their effectiveness declines across all artifact types, with text-based artifacts remaining the most resilient. This suggests that larger and more visible artifacts are more likely to be leveraged by the model, while smaller or faded artifacts are either ignored or contribute less to misclassification. Notably, the steepest decline occurs for graphics-based artifacts, particularly those without text, indicating that purely visual

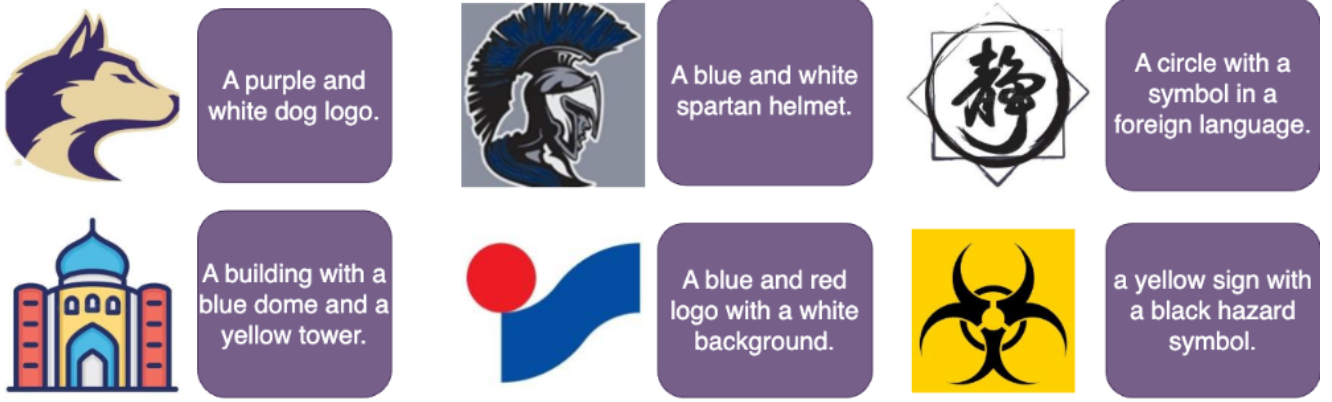


Figure 5. **Examples of generated captions for Web Artifact Attacks mitigation.** Each image represents a graphic artifact, accompanied by its corresponding caption generated by a vision-language model. These captions are then included in the prompt to mitigate the attack.

artifacts are more sensitive to reductions in size and visibility. These findings highlight that while reducing artifact saliency can mitigate their impact, text-based artifacts still pose a considerable threat, even when minimally visible.

### E. Descriptions of Artifacts for Mitigation

Fig. 5 illustrates how automatically generated captions are incorporated into our artifact-aware prompting strategy (see Sec 5) to mitigate Web Artifact Attacks. Inspired by prior work [1], which demonstrated that Vision-Language Models (VLMs) can adjust their attention when given more informative prompts, we extend this approach to graphical artifacts. Unlike text-based artifacts, which can be directly embedded into prompts, graphical artifacts lack an explicit textual representation, making them harder for the model to explicitly consider. To address this, we generate structured descriptions of graphical symbols and append them to the input prompt, ensuring that the model processes them explicitly rather than forming unintended associations.

The captions in Fig. 5 serve this purpose by neutralizing spurious correlations and guiding the model toward actual visual semantics. For example, instead of allowing the model to infer associations based on dataset biases (e.g., linking a hazard symbol to danger-related concepts), the caption describes it as “a yellow sign with a black hazard symbol”, removing any loaded interpretation.

### F. Beyond Visual Recognition

While our primary focus was on classification tasks for the sake of clarity and control, we also extend our attacks to the image retrieval task using the Flickr30K dataset [5]. On Top-1 Image-to-Text retrieval, our attacks achieve an 83.4% success rate on Graphics w/ text, 63.8% success rate on Graphics wo/ text, and 63.4% on Text, demonstrating their ability to generalize to applications beyond classification.

### References

- [1] Hao Cheng, Erjia Xiao, Jindong Gu, Le Yang, Jinhao Duan, Jize Zhang, Jiahang Cao, Kaidi Xu, and Renjing Xu. Unveiling typographic deceptions: Insights of the typographic vulnerability in large vision-language model. *ECCV*, 2024. 4
- [2] Kimmo Karkkainen and Jungseock Joo. Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1548–1558, 2021. 1
- [3] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3730–3738, 2015. 1
- [4] Subhansu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013. 1
- [5] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649, 2015. 4
- [6] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1