Adversarial Attacks on Image Generation With Made-Up Words

Raphaël Millière
Columbia University
`rm3799@columbia.edu`

Abstract

Text-guided image generation models can be prompted to generate images using nonce words adversarially designed to robustly evoke specific visual concepts. Two approaches for such generation are introduced: *macaronic prompting*, which involves designing cryptic hybrid words by concatenating subword units from different languages; and *evocative prompting*, which involves designing nonce words whose broad morphological features are similar enough to that of existing words to trigger robust visual associations. The two methods can also be combined to generate images associated with more specific visual concepts. The implications of these techniques for the circumvention of existing approaches to content moderation, and particularly the generation of offensive or harmful images, are discussed.

1 Introduction

Text-guided image generation models have made impressive strides in recent years. State-of-the-art models, like DALL-E 2 [1], Imagen [2], and Parti [3], can generate coherent images matching a remarkably wide variety of prompts in virtually any visual domain and style. While the ability to generate high-quality images of any subject is an exciting development for content creation, it also raises ethical questions about potential misuse of this technology. In particular, text-guided image generation models may be used to produce fake imagery of existing individuals for misinformation (so-called "deepfakes" [4]), or produce visual content deemed offensive or harmful. These concerns have been used to justify the decision to limit access to large text-guided image generation models, as well as moderate their use according to content policies implemented in prompt filters.

One particular concern regarding the robustness and apppropriate use of deep neural networks is the existence of so-called adversarial examples designed to mislead them. Adversarial examples have been extensively investigated in computer vision, where adding a small but carefully crafted perturbation to an image can lead image classifiers, but not humans, to make radical classification errors [5, 6]. The process of generating adversarial examples that can fool a given neural network is also known as an adversarial attack. In recent years, such attacks have been investigated beyond image classifiers. While adversarial perturbations can be indistinguable to the original input for humans in the visual domain, this is not typically the case in the linguistic domain, because text data is discrete rather than continuous like pixel values (although see [7] for an example of hidden adversarial attacks on text). Nonetheless, various alternative approaches for adversarial attacks have explored in natural language processing (NLP), causing models to misclassify text or generate

biased or harmful text with minimal perturbations of the input [8]. For example, appending a seemingly meaningless sequences of tokens to a paragraph can cause NLP models to fail at question answering tasks or spew racist outputs [9, 10].

Adversarial examples have also been explored with vision-language models designed for image captioning and recognition [11]. For example, so-called *typographic attacks* involve applying a real-life erroneous label to an item in an image (such as sticking a piece of paper with "ipod" written on it on an apple), causing vision-language models to misclassify that item [12]. However, adversarial attacks on text-guided image generation models have not been thoroughly investigated to date, despite common concerns about their potential misuse. Recent work suggests that it is possible to generate images corresponding to specific visual concepts by using seemingly nonsensical prompts with some text-guided image generation models. For example, Daras & Dimakis [13] suggested that gibberish text depicted in images generated by DALL-E 2 may belong to a "hidden vocabulary", which can be transcribed and used in prompts to generate specific kinds of images. Thus, they observed that the string `Apoploe vesrreaitais`, appearing in at least one image generated by the model, could be used as a prompt to yield images of birds. This method is potentially problematic, as it could be generalized to perform attacks on text-guided image generation models to get around content moderation filters. While this is not an adversarial attack in the traditional sense of introducing small or imperceptible perturbation to a model's input in order to drastically change its output, it does fall within the category of adversarial attacks in a broader sense: inputs are adversarially crafted to be meaningless and seemingly innocuous to humans, yet produce specific outputs associated with determinate visual concepts. If nonce strings can reliably be used to generate specific imagery with image generation models, then prompt filtering based on blacklisted words (e.g., violent, racist, sexist, or pornographic concepts) might turn out to be ineffective.

It is currently unclear how robust the association between gibberish strings and visual concepts is in text-guided image generation models. Many of the strings depicted in images generated by DALL-E 2 do not seem consistently associated with particular visual concepts. Even those that do, like `Apoploe vesrreaitais`, may not be associated to visual concepts with the same degree of robustness as real English words like `bird`. Furthermore, this phenomenon may be partially or wholly attributable to tokenization with *byte pair encoding* (BPE) [14, 15] used to train the CLIP model used for DALL-E 2 [1]. Thus, `Apoploe vesrreaitais` is tokenized with BPE as `Apoploe` `vesrreaitais`. Interestingly, *Apodidae* and *Ploceidae* are Latin names for bird families, and get tokenized by CLIP as `apodidae` and `ploceidae` respectively (alternating colors indicate subword segmentation). It is possible that the association between `Apoploe vesrreaitais` and images of bird is at least partially mediated by sharing initial tokens with `Apodidae` and `Ploceidae`. The claim that DALL-E 2 has developed a "hidden vocabulary" may be misleading, but highlights the importance of investigating the relationship between subword tokenization and image generation in text-guided image generation models, and, more broadly, viable strategies for adversarial attacks with nonce strings.

One significant limitation of the adversarial approach proposed by Daras & Dimakis [13] is that it does not offer a reliable method to find nonce strings that elicit specific imagery. Most of the gibberish text included by DALL-E 2 in images does not seem to be reliably associated with specific visual concepts when transcribed and used as a prompt. This limits the viability of this approach as way to circumvent the moderation of harmful or offensive content; as such, it is not a particularly concerning risk for the misuse of text-guided image generation models. The present work introduces two alternative methods for adversarial attacks on text-guided image generation models using nonce strings. *Macaronic prompting* involves composing multilingual subword units to generate specific

imagery from seemingly nonsensical strings, while *evocative prompting* relies on inventing words whose morphological features trigger consistent semantic associations with real concepts. Both methods can potentially be used to avoid triggering content filters, which pauses a potential risk for the safe deployment of text-guided image generation models. Macaronic prompting, in particular, appears to be significantly more reliable than the method proposed by Daras & Dimakis [13] to prompt such models with nonce strings. Furthermore, the allows the same adversarial attacks to work effectively on different models, instead of being tailored to one specific model like DALLE-2.

2 Macaronic prompting

The term "macaronic" traditionally refers to a mixture of several languages. One specific phenomenon that falls under this broad category is lexical hybridization, a form of code mixing in which words or morphemes from multiple languages are combined into novel words. Lexical hybridization spontaneously occurs in many multilingual communities. For example, Urdu-English code mixing is common in Pakistan, and includes the combination of English nouns with Urdu suffixes [16].

Text-guided image generation models using subword tokenization learn robust associations between subword units and visual concepts. Furthermore, they typically learn such associations in multiple languages, as training datasets scraped from the internet often contain image-caption pairs in different languages, even when multilinguality was not explicitly part of the curation goals. The core idea of macaronic prompting is to leverage such associations across different languages, in a way that purposefully obfuscates the nature of the visual concepts targeted in the prompt. This is achieved through a form of artificial lexical hybridization that involves combining multilingual subword units associated with similar visual concepts to produce nonce strings that do not trigger robust semantic associations in humans.

For example, the word for `birds` is `Vögel` in German, `uccelli` in Italian, `oiseaux` in French, and `pájaros` in Spanish. With the BPE tokenization method used by CLIP, these words are tokenized as follows: `vogel`, `uccelli`, `oiseaux`, and `pajaros`. Subword tokens from these words can be creatively combined to form a nonce words that looks like gibberish to human speakers, such as `uccoisegeljaros`.[1] Sample images generated by DALL-E 2 using this nonce string are shown in fig. 1.

Actual examples of code mixing in linguistic communities need not respect the so-called "equivalence constraint", according to which languages will tend be switched at points where their surface structures map onto each other [17]. By analogy, macaronic prompting need not strictly respect subword segmentation driven by BPE tokenization, and good results can in fact be obtained by carving out and recomposing words with arbitrary chunks. For example, `voiscellpajaraux` and `oisvogajaro` work just as well as `uccoisegeljaros` to generate images of birds (fig. 2).

Table 1 shows additional examples of lexical hybridization used for effective macaronic prompting about various visual concepts in different domains. Non-cherry-picked sample images generated with DALL-E 2 using the resulting nonce strings are displayed in fig. 3. These samples exhibit high consistency, despite the somewhat cryptic nature of the prompts for humans.

While different text-guided image generation models have different architectures, training data, and tokenization methods, macaronic prompting can in principle be applied to any model trained

---

[1] For clarity, the following color scheme is used to indicate the origin of each part of the nonce string: German, Italian, French, and Spanish.

Figure 1: Sample images generated by DALL-E 2 with the prompt `uccoisegeljaros`.

(a) Images generated by DALL-E 2 with the prompt `voiscellpajaraux`.

(b) Images generated by DALL-E 2 with the prompt `oisvogajaro`.

Figure 2: Examples of adversarial macaronic prompting that do not strictly respect the boundaries of subword tokenization for lexical hybridization.

English	German	Italian	French	Spanish	Hybridized (example)
bugs	Käfer	insetti	insectes	bichos	insekafetti
butterfly	Schmetterling	farfalla	papillon	mariposa	farpapmaripterling
lizard	Eidechse	lucertola	lézard	lagarto	eidelucertlagarzard
rabbit	Kaninchen	coniglio	lapin	conejo	coniglapkaninc
cliff	Klippe	scogliera	falaise	acantilado	falaiscoglieklippantilado
plane	Flugzeug	aereo	avion	avión	avflugzereo
firefighter	feuerwehrmann	pompiere	pompier	bombero	Feuerpompbomber
education	Bildung	educazione	éducation	educación	educbildacion
exasperation	Enttäuschung	esasperazione	exaspération	exasperación	exaspenttausacion

Table 1: Additional examples of hybridized made-up words used for macaronic prompting.

4

(a) Prompt: `insekafetti`.

(b) Prompt: `farpapmaripterling`.

(c) Prompt: `eidelucertlagarzard`.

(d) Prompt: `coniglapkaninc`.

(e) Prompt: `falaiscoglieklippantilado`.

(f) Prompt: `avflugzereo`.

(g) Prompt: `feuerpompbomber`.

(h) Prompt: `educbildacion`.

Figure 3: Additional examples of macaronic prompting with DALL-E 2.

on multilingual data. In fact, some macaronic prompts appear to work consistently across models, suggesting that it might be a viable approach for one-size-fits-all adversarial attacks on image generation algorithms. Fig. 4 shows some samples generated by DALL-E mini [18] with macaronic prompts that also work with DALL-E 2. It is worth noting that, despite their similar names, DALL-E 2 and DALL-E mini are fairly different. They have different architectures (notably, DALL-E mini does not use diffusion), are trained on different datasets, and use different tokenization procedures (DALL-E mini uses a BART tokenizer that may segment differently than CLIP's tokenizer). Given these differences, the fact that many macaronic prompts work on both models is noteworthy.

(a) Images generated by DALL-E mini with the prompt: `insekafetti`.

(b) Images generated by DALL-E mini with the prompt: `falaiscoglieklippantilado`.

(c) Images generated by DALL-E mini with the prompt: `avflugzereo`.

(d) Images generated by DALL-E mini with the prompt: `educbildacion`.

Figure 4: Examples of macaronic prompts that work across models.

However, not all macaronic prompts transfer appropriately across models. For example, while `farpapmaripterling` produces images of butterflies with DALL-E 2 as intended (fig. 3b), it unexpectedly produces images of mushrooms with DALL-E mini (fig. 5b). Interestingly, using the different hybridized nonce word `maripofarterling` yields images of butterflies in both DALL-E 2 and DALL-E mini (fig. 5b). Other examples of transfer failure include `voiscellpajaraux` and `eidelucertlagarzard`, which produce images of houses with DALL-E mini, instead of images of birds and lizards, respectively (fig. 5c,d). It is possible that larger models trained on larger datasets are more susceptible to macaronic prompting, as they learn more robust associations between subword units and visual concepts across languages. This might explain why some macaronic prompts that yield the expected results with DALL-E 2 do not work with DALL-E mini, while the reverse does not seem to hold. This trend is potentially worrisome, as it suggests that larger models may be more vulnerable to adversarial attacks using macaronic prompting.

In addition to their standalone use as prompts, macaronic nonce strings can be composed in sentences using English syntax, with some success. Fig. 6 shows some examples of compositional prompts involving one or several macaronic nonce strings. The fact that such strings can be combined to generate more specific and complex scenes opens up further possibilities for the generation of potentially problematic content using this approach. While complex macaronic prompts require English (or another natural language sufficiently well-represented in the training data) as a scaffold for syntactic structure, thus making them somewhat more interpretable than prompts using unique strings, the information conveyed to the model remains relatively well obfuscated. For example, it would probably be challenging for most people to guess what kind of scene will be generated

6

(a) Images generated by DALL-E mini with the prompt: `farpapmaripterling`.

(b) Images generated by DALL-E mini with the prompt: `maripofarterling`.

(c) Images generated by DALL-E mini with the prompt: `voiscellpajaraux`.

(d) Images generated by DALL-E mini with the prompt: `eidelucertlagarzard`.

Figure 5: Examples of transfer failure and success across models.

with the prompt `An eidelucertlagarzard eating a maripofarterling` without prior exposure to macaronic prompting and knowledge of the languages used for hybridization. Furthermore, compositionally complex prompts of this kind will not trigger content filters based on blacklists, despite their use of plain English words, as long as the censored concepts are adequately "encrypted" using the macaronic method.

(a) Images generated with the prompt: `A man in a state of exaspenttausacion`.

(b) Images generated with the prompt: `A farpapmaripterling lands on a feuerpompbomber`.

(c) Images generated with the prompt: `An eidelucertlagarzard maripofarterling. A fantastical hybrid creature, part eidelucertlagarzard and part maripofarterling, digital art`.

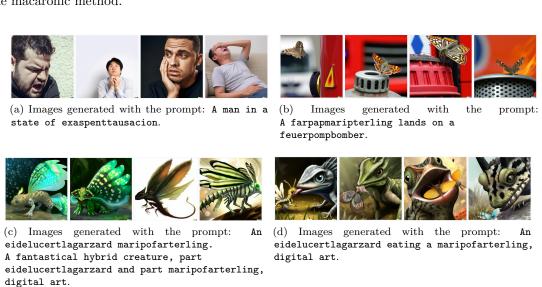(d) Images generated with the prompt: `An eidelucertlagarzard eating a maripofarterling, digital art`.

Figure 6: Examples of compositional macaronic prompting with DALL-E 2.

It is worth mentioning that lexical hybridization can also be effectively applied to prompting within a single language, although this will typically reveal more information about the queried visual concepts, and consequently be less relevant for adversarial attacks (i.e., it is likely that humans could guess the intended meaning of the nonce strings above chance). Fig. 7 shows

7

examples of images generated from multiple models using three invented English portmanteau words: `creepooky` (`creepy` + `spooky`), `happeerful` (`happy` + `cheerful`), and `lovssionate` (`loving` + `compassionate`).

(a) Images generated by DALL-E 2 with the prompt: `A very creepooky person.`

(b) Images generated by DALL-E mini with the prompt: `A very creepooky person.`

(c) Images generated by DALL-E 2 with the prompt: `A very happeerful person.`

(d) Images generated by DALL-E mini with the prompt: `A very happeerful person.`

(e) Images generated by DALL-E 2 with the prompt: `A very lovssionate person.`

(f) Images generated by DALL-E mini with the prompt: `A very lovssionate person.`

Figure 7: Examples of promtps using monolingual portmanteau words.

3 Evocative prompting

Macaronic prompting relies on existing morphemes or non-morphological word chunks to trigger visual associations in text-guided image generation models. By contrast, *evocative prompting* relies more loosely on broad morphological similarity with existing words, word categories, or language-specific features to trigger such associations. The key difference is that nonce strings used for evocative prompting need not include any part of an existing word in any language. Their association with specific visual concepts appears to be mediated instead by the statistical significance of certain letter combinations across words in a given domain.

A salient example of evocative prompting concerns the use of pseudolatin that presents superficial similarity to binomial nomenclature for biological taxonomy. This can be done to reliably prompt text-guided image generation models to generate images of the kinds of subjects typically designated through such nomenclature, namely organisms. Fig. 8 shows examples of evocative prompting with pseudolatin across models. Another domain in which evocative prompting can easily be achieved pertains to nonce strings that present a superficial similarity to medicine names. Fig. 9 displays samples produced with the prompts `vacyloraxin` and `walbotricypofen`.

Evocative prompting can also be applied to association of between language-specific features and visual features related to locations and cultures in which the corresponding languages are
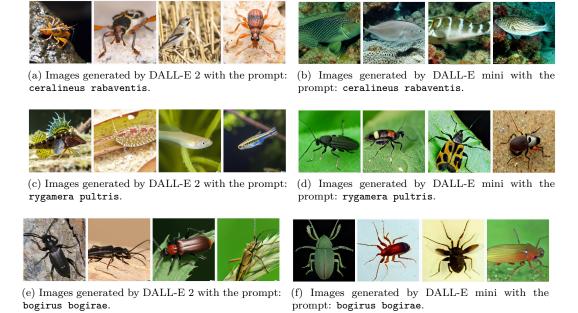
(a) Images generated by DALL-E 2 with the prompt: `ceralineus rabaventis`.

(b) Images generated by DALL-E mini with the prompt: `ceralineus rabaventis`.

(c) Images generated by DALL-E 2 with the prompt: `rygamera pultris`.

(d) Images generated by DALL-E mini with the prompt: `rygamera pultris`.

(e) Images generated by DALL-E 2 with the prompt: `bogirus bogirae`.

(f) Images generated by DALL-E mini with the prompt: `bogirus bogirae`.

Figure 8: Examples of evocative prompting with pseudolatin nomenclature.

(a) Images generated by DALL-E 2 with the prompt: `vacyloraxin`.

(b) Images generated by DALL-E 2 with the prompt: `walbotricypofen`.

Figure 9: Examples of evocative prompting with fictional medicine names.

spoken. Fig. 10 shows how prompting different models with made-up words that look like names of regional locations in different languages effectively generates images associated with statistically salient features of these locations. Thus, `Woldenbüchel` generates scenes that look like typical views of a German or Austrian village; `Valtorigiano` generates scenes that look like typical views of an old Italian town; and `Beaussoncour` causes DALL-E 2 to generate scenes that look like typical views of a historic French town. Interestingly, this last prompt causes DALL-E mini to generate images reminiscent of 17$^{\text{th}}$ century French portraits rather than French locations, but the association with French culture appears to be preserved nonetheless. Furthermore, the images produced by DALL-E mini with the prompts `voiscellpajaraux` and `eidelucertlagarzard` in fig. 5c-d could be seen as success cases of evocative prompting (French-looking strings associated with French-looking architecture) instead of failure cases of macaronic prompting. It is intriguing to consider where the boundary lies between macaronic and evocative prompting when real word chunks occur in the prompt; namely, in which cases the model is more sensitive to local features of the prompt (word chunks from any language present in the training data) as opposed to global features (overall morphological similarity to domain-specific words). It seems that differences in training data, model size, and model architecture may cause different models to parse prompts like `voiscellpajaraux` and `eidelucertlagarzard` in either "macaronic" or "evocative" fashion, even when these models are proven to be responsive to both prompting methods.

(a) Images generated by DALL-E 2 with the prompt: `Woldenbüchel`.

(b) Images generated by DALL-E mini with the prompt: `Woldenbüchel`.

(c) Images generated by DALL-E 2 with the prompt: `Valtorigiano`.

(d) Images generated by DALL-E mini with the prompt: `Valtorigiano`.

(e) Images generated by DALL-E 2 with the prompt: `Beaussoncour`.

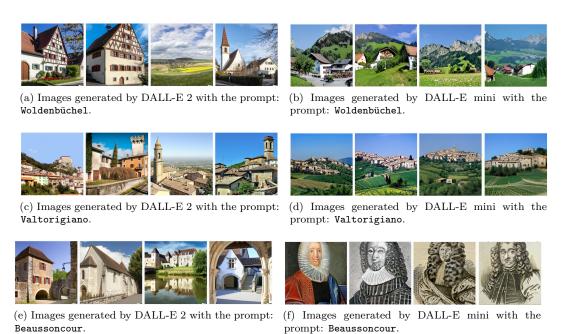(f) Images generated by DALL-E mini with the prompt: `Beaussoncour`.

Figure 10: Examples of evocative prompting with geographical associations.

Evocative prompting can also be combined with lexical hybridization to gain more control over the specific features of the outputs. For example, introduce English word chunks within pseudolatin nomenclature causes DALL-E 2 to generate images of animals with specific properties, as illustrated in fig. 11. Thus, the prompt `scariosus ferocianensis` (combining `scary` and `ferocious` with

pseudolatin terminations) yields images of conventionally scary "creepy crawlies" such as scorpions; `cutiosus adorablensis` (combining `cute` and `adorable` with pseudolatin terminations) yields images of conventionally cute mammals; `watosus swimensis` (combining `water` and `swimming` with pseudolatin terminations) yields images of aquatic animals; and `flyosus wingensis` (combining `flying` and `winged` with pseudolatin terminations) yields images of flying insects.
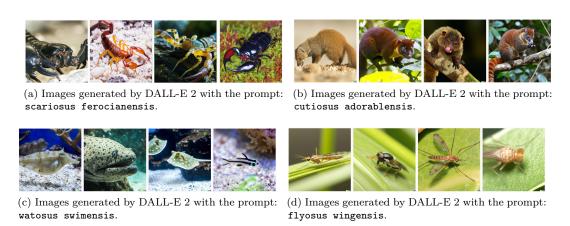

(a) Images generated by DALL-E 2 with the prompt: `scariosus ferocianensis`.

(b) Images generated by DALL-E 2 with the prompt: `cutiosus adorablensis`.

(c) Images generated by DALL-E 2 with the prompt: `watosus swimensis`.

(d) Images generated by DALL-E 2 with the prompt: `flyosus wingensis`.

Figure 11: Examples of evocative prompting combined with lexical hybridization.

4   Discussion

Vulnerabilty to various kinds of adversarial attacks is a significant limitation of many deep artificial neural networks. First, it highlights ways in which the behavior of these networks may be not be human-like, which has implications for their use as models of human cognition. Second, and more worryingly, adversarial attacks can be intentionally and maliciously deployed to trick neural networks into misclassifying inputs or generating problematic outputs, which may have real-life adverse consequences.

The new methods presented in this work, *macaronic promping* and *evocative prompting*, enable targeted adversarial attacks on text-guided image generation models. Macaronic prompting involves combining word chunks from different languages into novel hybridized words that lack meaning for human speakers, yet consistently trigger specific visual associations in image generation models. This approach is somewhat related to previous methods using code-mixing for adversarial attacks on multilingual NLP models [19]. The preliminary experiments shown here suggest that hybridized nonce strings can be methodically crafted to generate images of virtually any subject as needed, and even combined together to generate more complex scenes with some success. An obvious concern with this method is the circumvention of content filters based on blacklisted prompts. In principle, macaronic prompting could provide an easy and seemingly reliable way to bypass such filters in order to generate harmful, offensive, illegal, or otherwise sensitive content, including violent, hateful, racist, sexist, or pornographic images, and perhaps images infringing on intellectual property or depicting real individuals. Companies that offer image generation as a service have put a great deal of care into preventing the generation of such outputs in accordance with their content policy.

Consequently, macaronic prompting should be systematically investigated as a threat to the safety protocols 1used for commercial image generation.

Evocative prompting presents a less obvious threat, because it does not offer a principled way of crafting nonce strings that can effectively and reliably trigger specific visual associations. As such, it is mostly limited to vague associations with concepts linked to broad morphological features of words or languages. The present work investigates a few domains in which evocative prompting appears to be somewhat reliable and effective as a proof of concept (biological nomenclature, medicine names, and language-specific geographical/cultural names). It is likely that evocative prompting can be applied to many more subjects whose associated lexicon exhibits salient morphological similarities in various languages. It is worth noting that the combination of evocative prompting with real word chunks can target more specific visual associations (fig. 11); however, the resulting prompts are easily interpretable for humans, and may be easier to filter through blacklists.

Overall, maraconic prompting is more viable than evocative prompting as a method for malicious adversarial attacks on text-guided image generation models. It highlights the insufficiency of keyword-based blacklists for content filtration in such models. One way to minimize opportunities for the generation of problematic content is to curate datasets appropriately to exclude specific categories of images or image-caption pairs. For example, some large multimodal datasets on which image generation models are trained are known to contain harmful stereotypes [20]. However, dataset curation can only go so far in preventing the downstream generation of problematic content. A well-trained model with the right architecture should in principle be able to construct scenes depicting harmful or offensive imagery by combining individually innocuous concepts present in the training data. A potential complement to prompt filtering through blacklists is the use of a dedicated image classifier to filter visual outputs instead of text inputs. Alternatively, it may be sufficient to feed visual outputs back to the image generation model, retrieve related linguistic concepts in the joint vision-language embedding space, and apply text-based blacklists to the retrieved concepts. Such filtering might prove effective in preventing malicious image generation through text-based adversarial attacks. Another, more radical defense against such attacks would involve filtering all prompts whose words do not occur in a comprehensive multilingual vocabulary list, which would effectively prevent the use of nonce strings across the board. However, this would impose significant limitations of the use of image generation models, particularly when it comes to prompts containing proper names or resilience to typographical errors.

Unlike adversarial perturbations traditionally used in computer vision, attacks from macaronic or evocative prompting are not completely inscrutable to humans [21]. With enough knowledge of languages and morphological features, it should be possible to recover the intended visual associations of nonce strings used in these approaches. While this work has only investigated macaronic hybridized strings using five languages (English, French, German, Italian, and Spanish), it should be possible to make these nonce strings even more cryptic by adding word chunks from additional lesser-known languages, especially as image generation models get trained on increasingly multilingual datasets. There is another respect in which these techniques can remain somewhat inscrutable, in that it seems currently difficult to predict whether any specific model will respond to a given macaronic or evocative prompt as intended. This is apparent, for example, from the fact that DALL-E mini is vulnerable to some but not all macaronic prompts that prove effective with DALL-E 2 (fig. 5), or that it has different visual associations with some but not all evocative prompts (fig. 10). It is currently challenging to disentangle the factors that determine convergent and divergent behavior with macaronic and evocative prompting across models.

5   Conclusion

Text-guided image generation models are not immune to adversarial attacks. While various properties of these models – including size, architecture, tokenization prodecure and training data – may influence their vulnerability to text-based adversarial attacks, preliminary evidence discussed in this work suggests that some of these attacks may nonetheless work somewhat reliably across models. Further research is needed to determine the factors that determine how different models respond to macaronic and evocative prompting, as well as effective strategies to mitigate malicious use of these techniques for the generation of harmful, offensive, or otherwise problematic visual content.

References

[1] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical Text-Conditional Image Generation with CLIP Latents, April 2022.

[2] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding, May 2022.

[3] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, Ben Hutchinson, Wei Han, Zarana Parekh, Xin Li, Han Zhang, Jason Baldridge, and Yonghui Wu. Scaling Autoregressive Models for Content-Rich Text-to-Image Generation, June 2022.

[4] Raphaël Millière. Deep learning and synthetic media. *Synthese*, 200(3):231, May 2022. ISSN 1573-0964. doi: 10.1007/s11229-022-03739-2.

[5] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks, February 2014.

[6] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and Harnessing Adversarial Examples, March 2015.

[7] Nicholas Boucher, Ilia Shumailov, Ross Anderson, and Nicolas Papernot. Bad Characters: Imperceptible NLP Attacks. In *2022 IEEE Symposium on Security and Privacy (SP)*, pages 1987–2004, May 2022. doi: 10.1109/SP46214.2022.9833641.

[8] Wei Emma Zhang, Quan Z. Sheng, Ahoud Alhazmi, and Chenliang Li. Adversarial Attacks on Deep-learning Models in Natural Language Processing: A Survey. *ACM Transactions on Intelligent Systems and Technology*, 11(3):24:1–24:41, April 2020. ISSN 2157-6904. doi: 10.1145/3374217.

[9] Robin Jia and Percy Liang. Adversarial Examples for Evaluating Reading Comprehension Systems, July 2017.

[10] Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh. Universal Adversarial Triggers for Attacking and Analyzing NLP, January 2021.

[11] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision. *arXiv:2103.00020 [cs]*, February 2021.

[12] Gabriel Goh, Nick Cammarata †, Chelsea Voss †, Shan Carter, Michael Petrov, Ludwig Schubert, Alec Radford, and Chris Olah. Multimodal neurons in artificial neural networks. *Distill*, 2021. doi: 10.23915/distill.00030.

[13] Giannis Daras and Alexandros G. Dimakis. Discovering the Hidden Vocabulary of DALLE-2, May 2022.

[14] Philip Gage. A new algorithm for data compression. *C Users J.*, 12(2):23–38, February 1994. ISSN 0898-9788.

[15] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural Machine Translation of Rare Words with Subword Units, June 2016.

[16] Dr Sarwet Rasul (Sarwet Dilshad). Verb Hybridization: An Interesting Aspect of the Spread of English in Pakistan. *Journal of Research (Languages and Islamic Studies)*, 11(1):37–52, June 2007.

[17] Shana Poplack. Sometimes I'll start a sentence in Spanish Y TERMINO EN ESPAÑOL: Toward a typology of code-switching1. 18(7-8):581–618, January 1980. ISSN 1613-396X. doi: 10.1515/ling.1980.18.7-8.581.

[18] Boris Dayma, Suraj Patil, Pedro Cuenca, Khalid Saifullah, Tanishq Abraham, Phuc Le Khac, Luke Melas, and Ritobrata Ghosh. DALL·E Mini, July 2021.

[19] Samson Tan and Shafiq Joty. Code-Mixing on Sesame Street: Dawn of the Adversarial Polyglots, June 2021.

[20] Abeba Birhane, Vinay Uday Prabhu, and Emmanuel Kahembwe. Multimodal datasets: Misogyny, pornography, and malignant stereotypes, October 2021.

[21] Cameron Buckner. Understanding adversarial examples requires a theory of artefacts for deep learning. *Nature Machine Intelligence*, 2(12):731–736, December 2020. ISSN 2522-5839. doi: 10.1038/s42256-020-00266-y.