THE UNIVERSITY OF SYDNEY

NEURAL INFORMATION PROCESSING SYSTEMS

Stable Diffusion is Unstable

Thirty-seventh Conference on Neural Information Processing Systems

Poster Name: Chengbin Du

Supervisor: Dr. Chang Xu

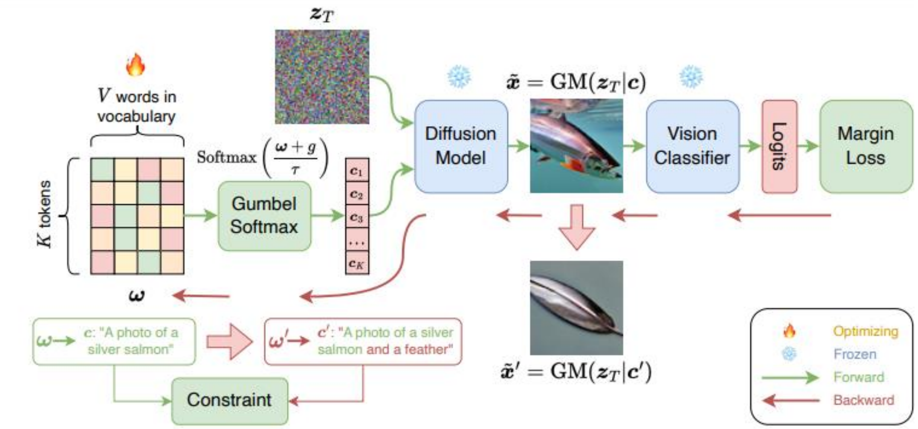
Presentation

Stable diffusion is Unstable

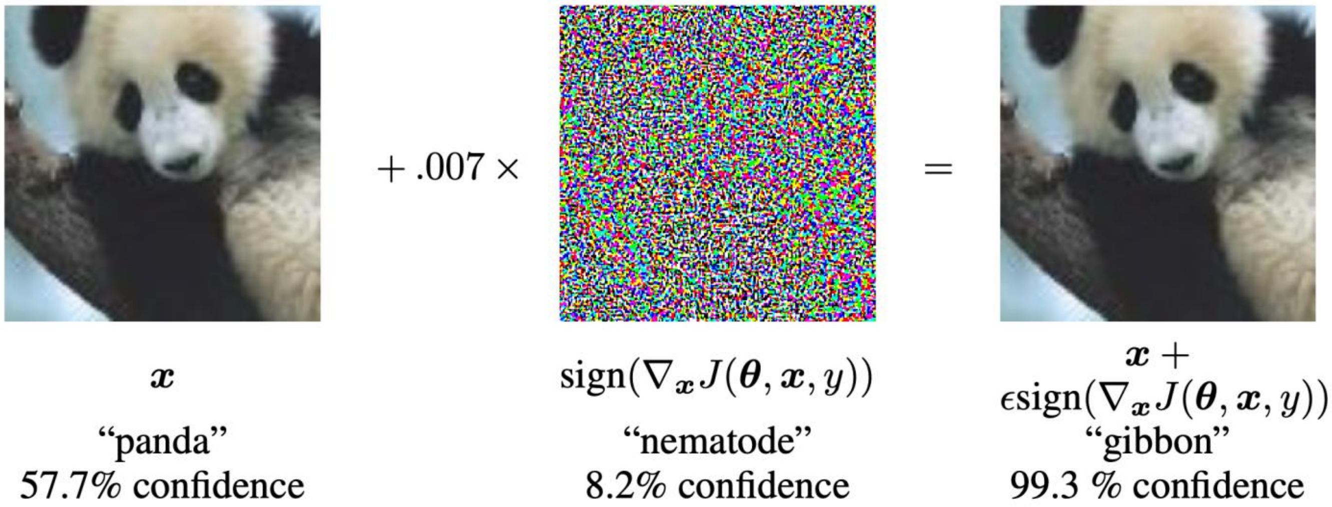
Chengbin Du, Yanxi Li, Zhongwei Qiu, Chang Xu

Abstract

Recently, text-to-image models have been thriving. Despite their powerful generative capacity, our research has uncovered a lack of robustness in this generation process. Specifically, the introduction of small perturbations to the text prompts can result in the blending of primary subjects with other categories or their complete disappearance in the generated images. In this paper, we propose Auto-attack on Text-to-image Models (ATM), a gradient-based approach, to effectively and efficiently generate such perturbations. By learning a Gumbel Softmax distribution, we can make the discrete process of word replacement or extension continuous, thus ensuring the differentiability of the perturbation generation. Once the distribution is learned, ATM can sample multiple attack samples simultaneously. These attack samples can prevent the generative model from generating the desired subjects without tampering with the category keywords in the prompt. ATM has achieved a 91.1% success rate in short-text attacks and an 81.2% success rate in long-text attacks. Further empirical analysis revealed four attack patterns based on: 1) the variability in generation speed, 2) the similarity of coarse-grained characteristics, and 3) the polysemy of words. The code is available at https://github.com/duchengbin8/Stable_Diffusion_is_Unstable


z_T
V words in vocabulary
K tokens
Softmax $\left(\frac{\omega + g}{\tau}\right)$
Gumbel Softmax
c_1
c_2
c_3
$\dots$
c_K
ω
Diffusion Model
$\tilde{x} = \mathrm{GM}(z_T|c)$
Vision Classifier
Logits
Margin Loss
$\tilde{x}' = \mathrm{GM}(z_T|c')$
$\omega \rightarrow$ c: "A photo of a silver salmon"
$\omega' \rightarrow$ c': "A photo of a silver salmon and a feather"
Constraint
Optimizing
Frozen
Forward
Backward

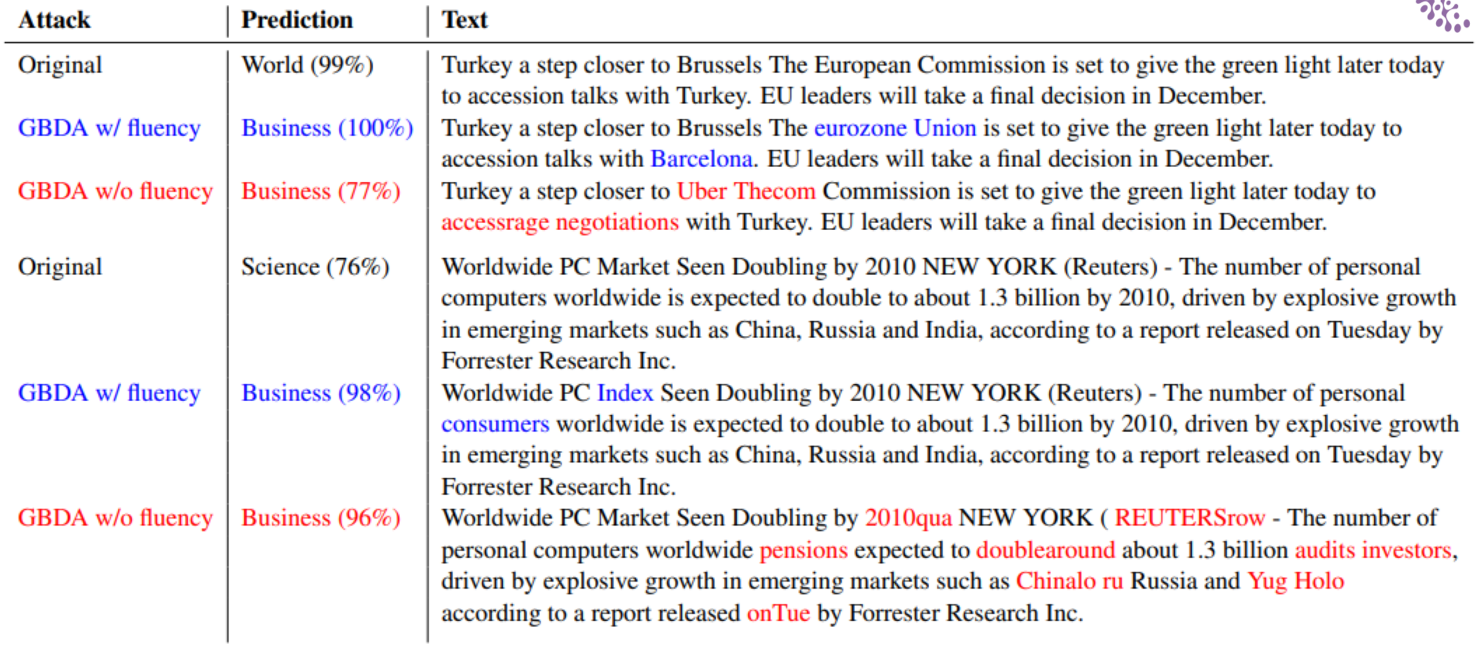
Background: Traditional Adversarial Attack


$+ .007 \times$
$=$
$\boldsymbol{x}$
"panda"
57.7% confidence
$\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$
"nematode"
8.2% confidence
$\boldsymbol{x} +$
$\epsilon\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$
"gibbon"
99.3 % confidence

An example of adversarial attacks on image: Fast Gradient Sign Method (FGSM)

Figure source: *Explaining and Harnessing Adversarial Examples*, Goodfellow *et al.*, In *ICLR*, 2015.

Background: Traditional Adversarial Attack



Attack	Prediction	Text
Original	World (99%)	Turkey a step closer to Brussels The European Commission is set to give the green light later today to accession talks with Turkey. EU leaders will take a final decision in December.
GBDA w/ fluency	Business (100%)	Turkey a step closer to Brussels The eurozone Union is set to give the green light later today to accession talks with Barcelona. EU leaders will take a final decision in December.
GBDA w/o fluency	Business (77%)	Turkey a step closer to Uber Thecom Commission is set to give the green light later today to accessrage negotiations with Turkey. EU leaders will take a final decision in December.
Original	Science (76%)	Worldwide PC Market Seen Doubling by 2010 NEW YORK (Reuters) - The number of personal computers worldwide is expected to double to about 1.3 billion by 2010, driven by explosive growth in emerging markets such as China, Russia and India, according to a report released on Tuesday by Forrester Research Inc.
GBDA w/ fluency	Business (98%)	Worldwide PC Index Seen Doubling by 2010 NEW YORK (Reuters) - The number of personal consumers worldwide is expected to double to about 1.3 billion by 2010, driven by explosive growth in emerging markets such as China, Russia and India, according to a report released on Tuesday by Forrester Research Inc.
GBDA w/o fluency	Business (96%)	Worldwide PC Market Seen Doubling by 2010qua NEW YORK (REUTERSrow - The number of personal computers worldwide pensions expected to doublearound about 1.3 billion audits investors, driven by explosive growth in emerging markets such as Chinalo ru Russia and Yug Holo according to a report released onTue by Forrester Research Inc.

An example of adversarial attacks on text: Gradient-based Adversarial Attacks against Text Transformers

Background: Stable diffusion model

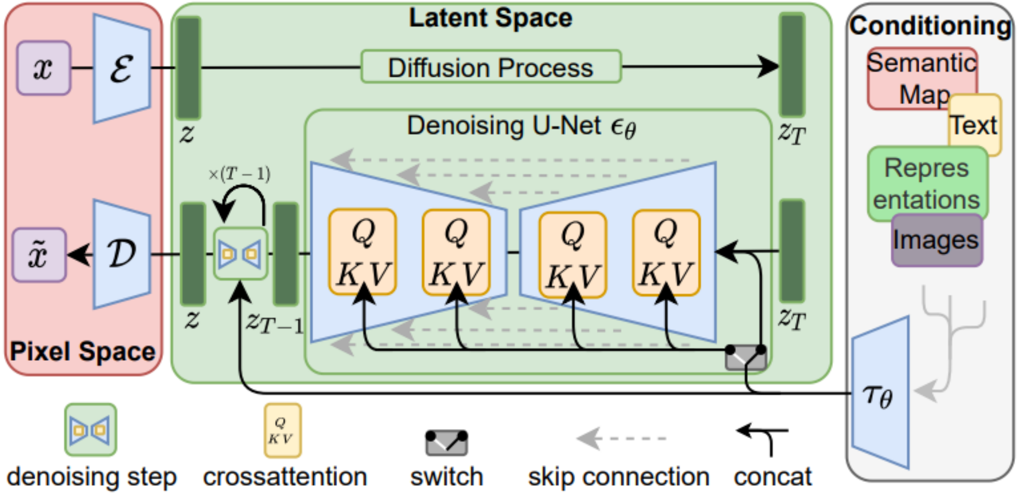


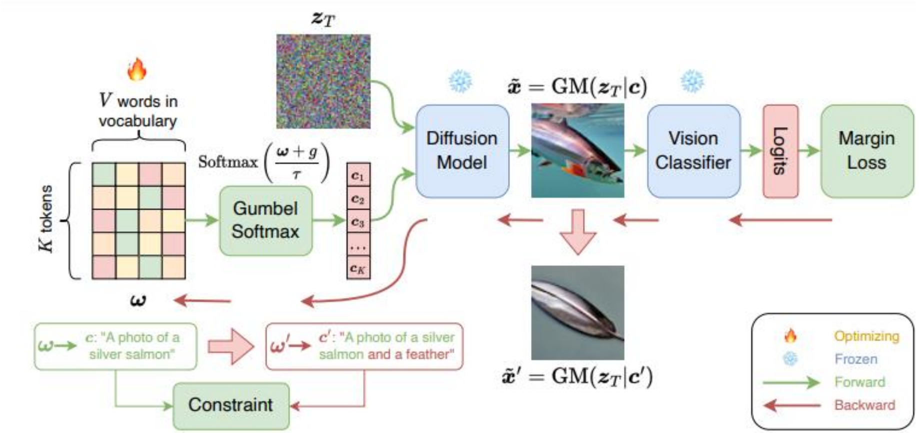
Figure 3. We condition LDMs either via concatenation or by a more general cross-attention mechanism. See Sec. 3.3

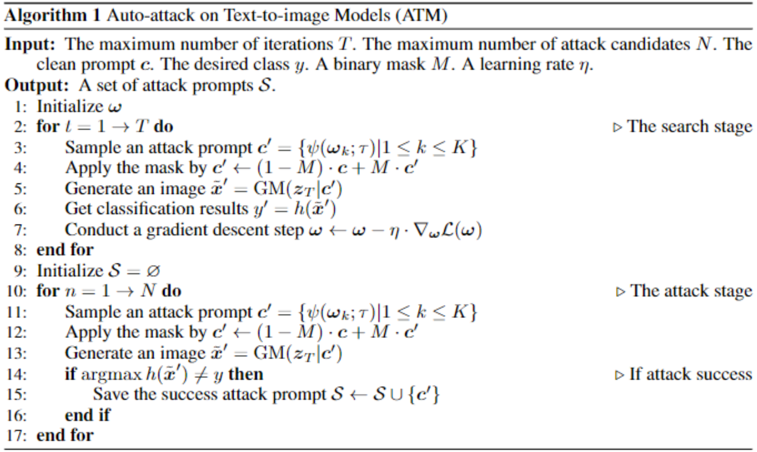
How can we extend the idea of adversarial attacks to text-to-image generative models?

The mechanism of adversarial attacks.

Figure source: *Geometry-aware Instance-reweighted Adversarial Training*, Zhang *et al.*, In *ICLR*, 2021.

An Adversarial Attack Paradigm for Text-to-Image Generative Models


z_T
V words in vocabulary
K tokens
ω
Softmax $\left(\frac{\omega+g}{\tau}\right)$
Gumbel Softmax
c_1
c_2
c_3
...
c_K
Diffusion Model
$\tilde{x} = \mathrm{GM}(z_T|c)$
Vision Classifier
Logits
Margin Loss
$\tilde{x}' = \mathrm{GM}(z_T|c')$
$\omega \rightarrow$ c: "A photo of a silver salmon"
$\omega' \rightarrow$ c': "A photo of a silver salmon and a feather"
Constraint
🔥 Optimizing
❄ Frozen
Forward
Backward



Algorithm 1 Auto-attack on Text-to-image Models (ATM)

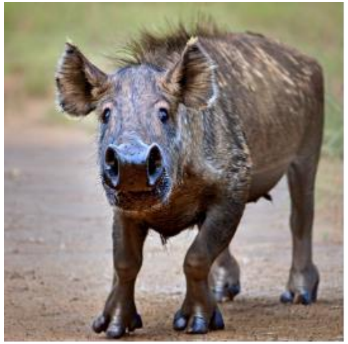
Input: The maximum number of iterations T. The maximum number of attack candidates N. The clean prompt c. The desired class y. A binary mask M. A learning rate η.

Output: A set of attack prompts $\mathcal{S}$.

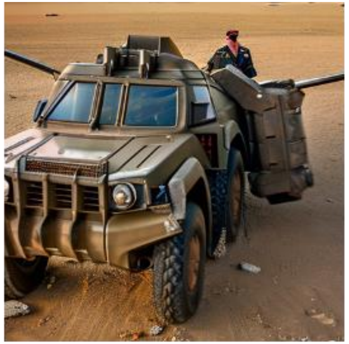
```
Initialize $\omega$
for $t = 1 \rightarrow T$ do                                          ▷ The search stage
    Sample an attack prompt $c' = \{\psi(\omega_k; \tau) | 1 \leq k \leq K\}$
    Apply the mask by $c' \leftarrow (1 - M) \cdot c + M \cdot c'$
    Generate an image $\tilde{x}' = \mathrm{GM}(z_T | c')$
    Get classification results $y' = h(\tilde{x}')$
    Conduct a gradient descent step $\omega \leftarrow \omega - \eta \cdot \nabla_\omega \mathcal{L}(\omega)$
end for
Initialize $\mathcal{S} = \varnothing$
for $n = 1 \rightarrow N$ do                                          ▷ The attack stage
    Sample an attack prompt $c' = \{\psi(\omega_k; \tau) | 1 \leq k \leq K\}$
    Apply the mask by $c' \leftarrow (1 - M) \cdot c + M \cdot c'$
    Generate an image $\tilde{x}' = \mathrm{GM}(z_T | c')$
    if $\arg\max h(\tilde{x}') \neq y$ then                            ▷ If attack success
        Save the success attack prompt $\mathcal{S} \leftarrow \mathcal{S} \cup \{c'\}$
    end if
end for
```

Cases

When given a text prompt: "A photo of a warthog", the Stable Diffusion model can generate a corresponding, clear and realistic photo of a warthog. However, when we slightly modify this text prompt to become: "A photo of a warthog and a traitor", the generated image becomes something that has nothing to do with the animal warthog at all.



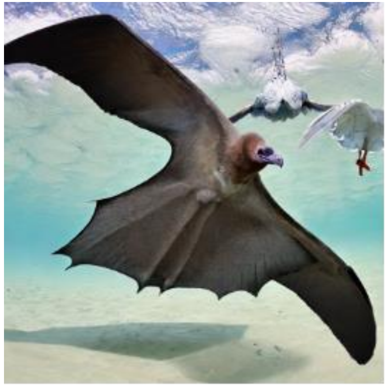
A photo of a warthog



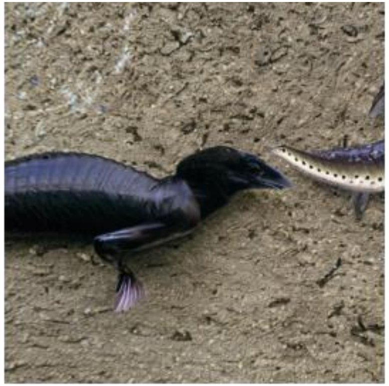
A photo of a warthog and a traitor

Another Cases

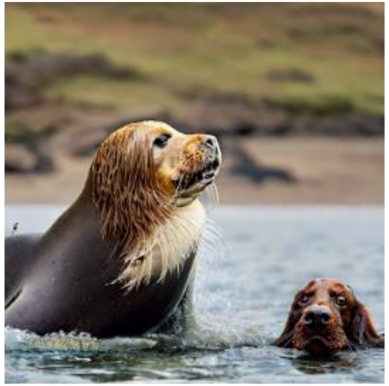
Stable diffusion model create new species



A photo of a
stingray and
a vulture



a photo of a
eel and
a rook flew



A photo of
a sea lion and
a spaniel

Pattern1: Variability in Generation Speed

In this Pattern, it is explained that when two objects need to be generated at the same time, the object whose generation speed is slow will eventually fail to appear in the generated image.

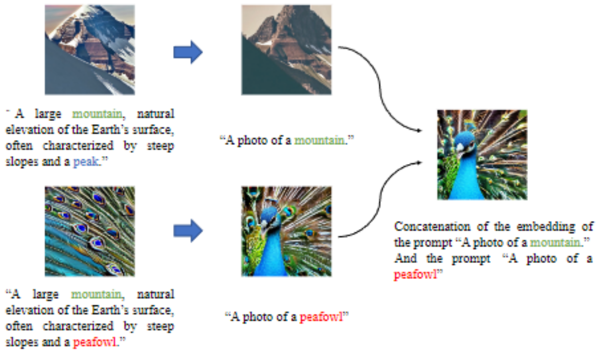

" A large mountain, natural elevation of the Earth's surface, often characterized by steep slopes and a peak."
"A photo of a mountain."
"A large mountain, natural elevation of the Earth's surface, often characterized by steep slopes and a peafowl."
"A photo of a peafowl"
Concatenation of the embedding of the prompt "A photo of a mountain." And the prompt "A photo of a peafowl"

Figure 2: A case study on "mountain" and "peafowl".

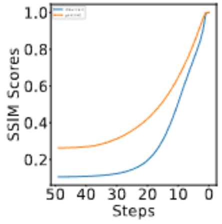


Figure 3: The generation speeds of "mountain" and "peafowl".

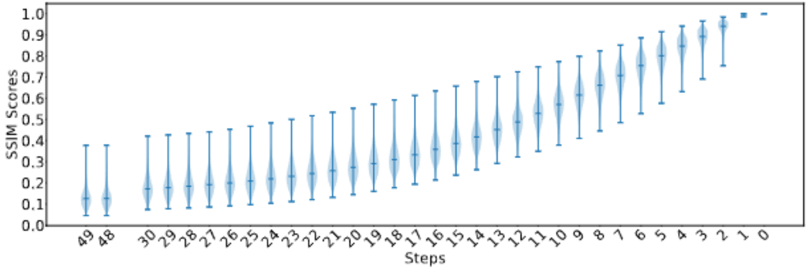


Figure 4: A violin plot illustrating the generation speeds of 1,000 images of various classes. The horizontal axis represents the number of steps taken, ranging from 49 to 0, while the vertical axis displays the SSIM scores. The width of each violin represents the number of samples that attained a specific range of SSIM scores at a given step.

Pattern2: Similarity of Coarse-grained Characteristics

In this Pattern, it is explained that when two objects need to have a certain coarse-grained similarity, they will undergo feature entanglement with a certain probability.

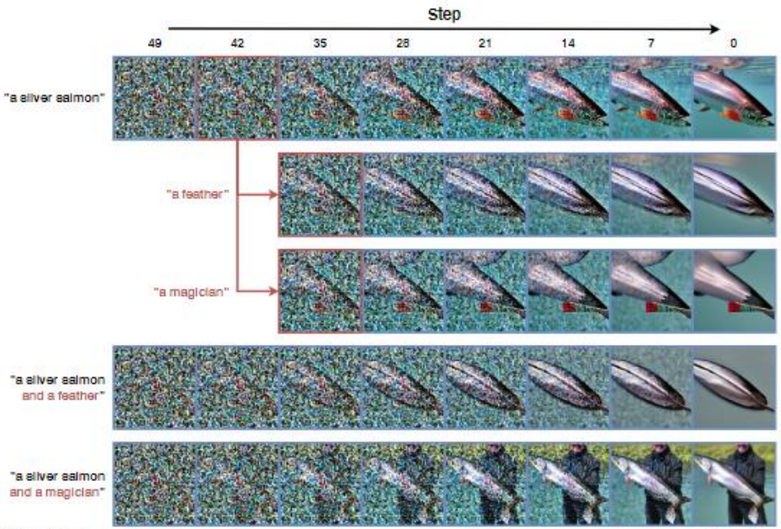

Step
49
42
35
28
21
14
7
0
"a silver salmon"
"a feather"
"a magician"
"a silver salmon and a feather"
"a silver salmon and a magician"

Figure 6: The first row illustrates the generation process with the prompt "a photo of a silver salmon". The second row, based on the forty-second step of the first row, shows the generation process with the prompt "a photo of a feather". The third row, also building upon the forty-second step of the first row, presents the generation procedure when the prompt is "a photo of a magician". The fourth row depicts the generation process in the presence of feature entanglement. The fifth row demonstrates the generation process for two distinct categories without feature entanglement.

Pattern3: Polysemy of Words

In this Pattern, it is explained that when a word has multiple meanings, adding perturbations can cause the word to deviate from the original semantics, thus affecting the final generated image.

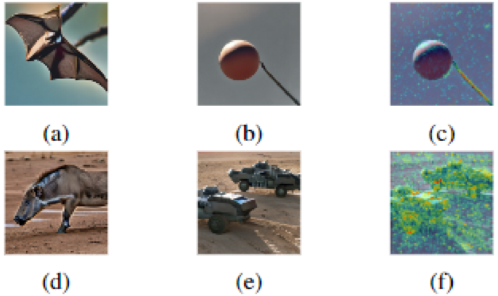
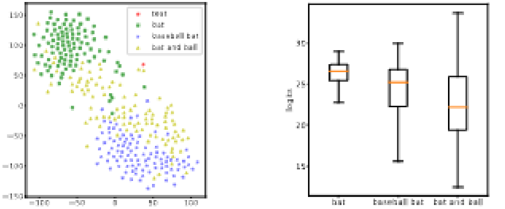


Figure 7: a) "A photo of a bat"; b) "A photo of a bat and a ball;" c) Heat map of the word "bat" in generated image; d) "A photo of a warthog"; e) "A photo of a warthog and a traitor"; f) Heat map of the word "warthog" in generated image.



(a) tsne

(b) logits

Figure 8: a) t-SNE Visualization of 100 images each of "bat", "baseball bat", "bat and ball“ and text "a photo of a bat." b) The boxplot of cosine similarities between the text embedding of "a photo of a bat" and 100 of image embeddings each of "bat", "baseball bat", and "bat and ball".

Thank you!

NEURAL INFORMATION PROCESSING SYSTEMS

THE UNIVERSITY OF SYDNEY