

Prompt Stealing Attacks Against Text-to-Image Generation Models

Xinyue Shen, Yiting Qu, Michael Backes, Yang Zhang

CISPA Helmholtz Center for Information Security





Art Made With A.I. Won a State Fair Last Year. Now, the Rules Are Changing



Painting

Square Enix says it used AI art in upcoming Foamstars game



Game

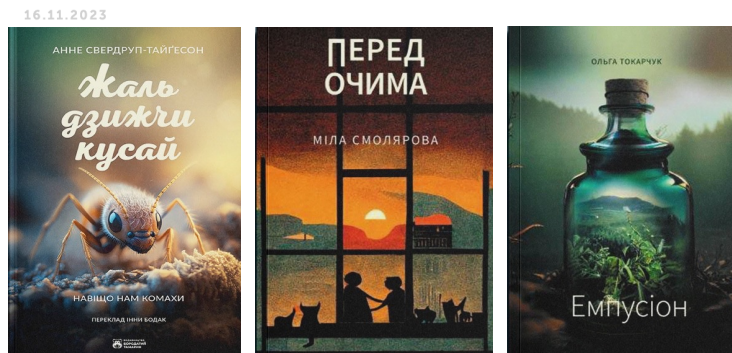
The World's Smartest Artificial Intelligence Just Made Its First Magazine Cover

The technology behind DALL-E 2 is already reshaping the world as you know it—perhaps most literally with this magazine cover you're looking at. Are you ready for what comes next?



Magazine Cover

Ukraine's book industry debates the rise of AI in publishing, balancing 'innovation' with ethical concerns



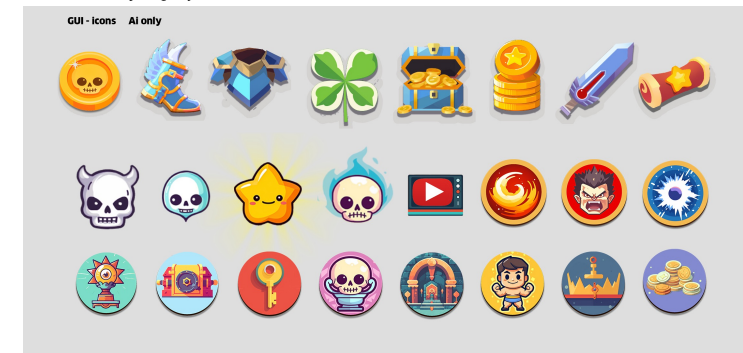
Book Illustration

Game Development 31.07.2023 16:15

in f t

Pros and cons of using Midjourney in game development — lessons learned by Nexters

Published by Evgeny Obedkov

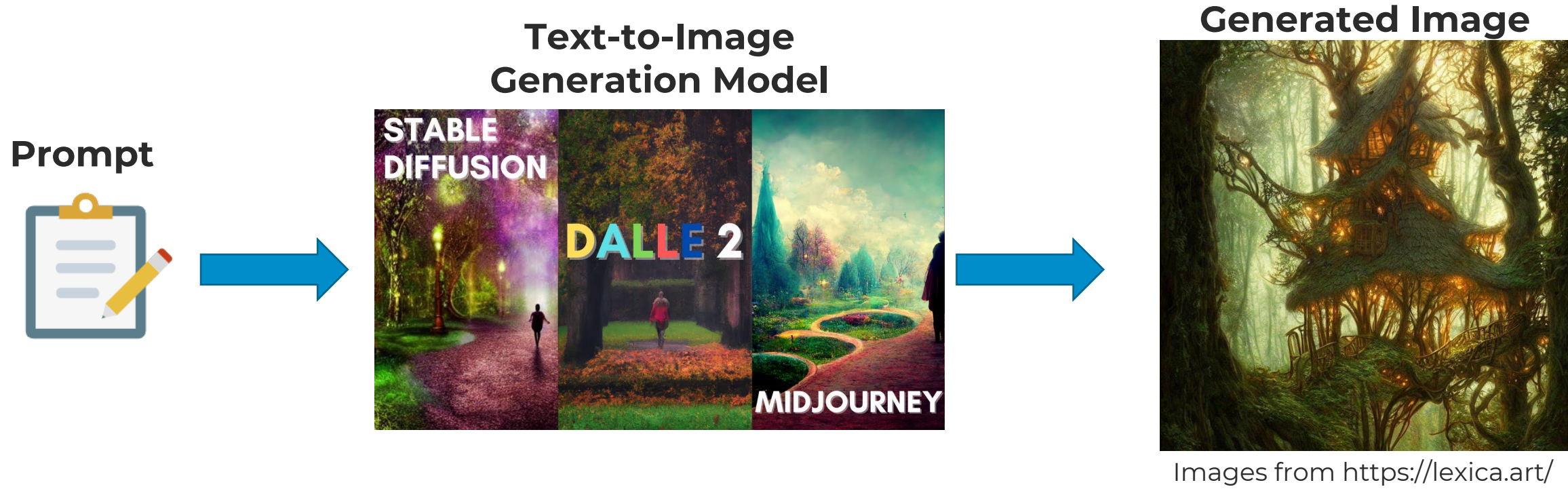


Game Assets

- [1] <https://www.smithsonianmag.com/smart-news/this-state-fair-changed-its-rules-after-a-piece-made-with-ai-won-last-year-180982867/>
- [2] <https://www.theverge.com/2024/1/16/24040124/square-enix-foamstars-ai-art-midjourney>
- [3] <https://www.cosmopolitan.com/lifestyle/a40314356/dall-e-2-artificial-intelligence-cover/>
- [4] <https://chytomo.com/en/ai-in-book-design-a-valuable-tool-or-underrating-of-work/>
- [5] <https://promptbase.com/prompt-purchase/4PPwdxOciCexceN48e62>



What is Text-to-Image Generation Model





Prompt for Text-to-Image Generation Model



Generated Image



Prompt

A treehouse in ancient forest



Prompt for Text-to-Image Generation Model

Prompt

A treehouse in ancient forest, diffuse lighting, fantasy, intricate, elegant, highly detailed, lifelike, photorealistic, digital painting, artstation, illustration, concept art, smooth, sharp focus, art by John Collier and Albert Aublet and Krenz Cushart and Artem Demura and Alphonse Mucha.

Subject depicts the main object

Generated Image

Text-to-Image Generation Model



Based on our measurement on 61K real-world prompts

Modifiers describe styles

- **Trending:** Pixiv, Pinterest, ...
- **Artist:** wlop, greg rutkowski, artgerm,
- **Medium:** concept art, digital art, ...
- **Movement:** fantasy art, photorealism, ...
- **Flavor:** highly detailed, 8k, ...



A story of prompt engineer @anthony



Prompt Engineer



Ideal Image
In @anthony's brain



A story of prompt engineer @anthony



Prompt Engineer

A mermaid

Prompt



Obtained Image



Ideal Image
In @anthony's brain



A story of prompt engineer @anthony



Prompt Engineer

Watercolor illustration of a mermaid. Ethereal and mystical atmosphere. Ancient architecture. Smoky and misty effects.

Prompt



Obtained Image



Ideal Image
In @anthony's brain



A story of prompt engineer @anthony



Prompt Engineer

Watercolor illustration of a mermaid. Ethereal and mystical atmosphere
...
soft, diffused lighting
...
Delicate brush strokes
...
nature and ancient architecture
....
Smoky and misty effects
...
enchanting mood.

Prompt



Obtained Image



Ideal Image
In @anthony's brain



@anthony is proud of his prompt, so he sells his prompt on a prompt marketplace

PromptBase Search prompts Marketplace Apps Create Sell Login

Models Art Logos Graphics Productivity Marketing Photography Games

Midjourney

Misty Watercolor Illustrations

2 Favorites | 11 Views

54 words V6.0 Tested Tips 9 examples HQ images

@anthony Top Seller 6.0k 4.6 ★★★★★ 276 ratings

Immerse yourself in ethereal watercolor illustrations that blend fantasy with nature and ancient architecture. Soft lighting, intricate details, and smoky effects create a surreal, enchanting mood. Perfect for adding a dreamlike, mystical touch to your creative projects. [...more](#)

\$2.99

Get prompt

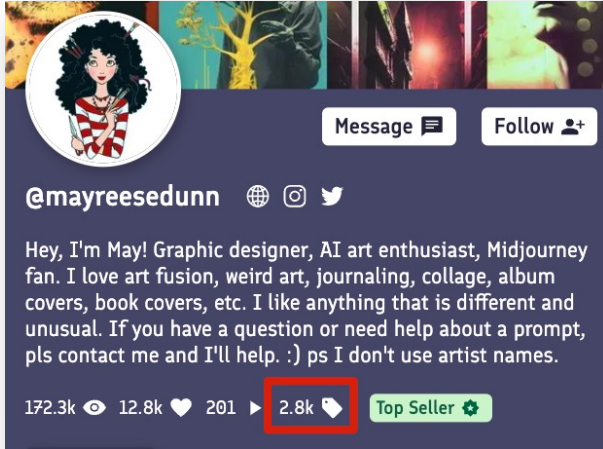
After purchasing, you will gain access to the prompt file which you can use with Midjourney. You'll receive 20 free generation credits with this purchase. By purchasing this prompt, you agree to our [terms of service](#).

3 weeks ago

These prompts bring @authony \$ 18,000+!



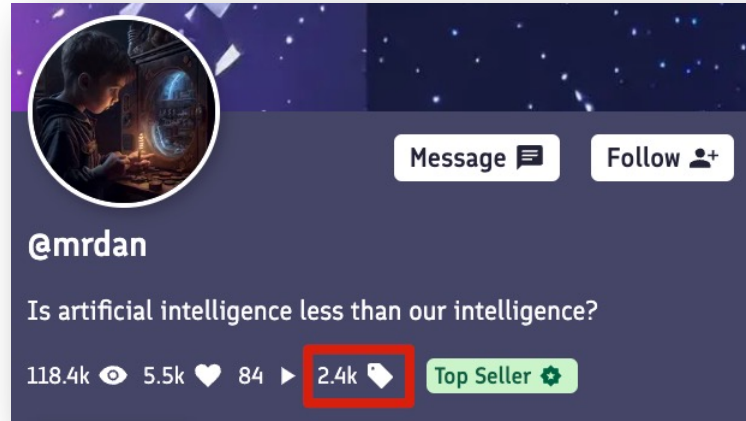
There are thousands of prompt engineers like @anthony



@mayreesedunn Message Follow

Hey, I'm May! Graphic designer, AI art enthusiast, Midjourney fan. I love art fusion, weird art, journaling, collage, album covers, book covers, etc. I like anything that is different and unusual. If you have a question or need help about a prompt, pls contact me and I'll help. :) ps I don't use artist names.

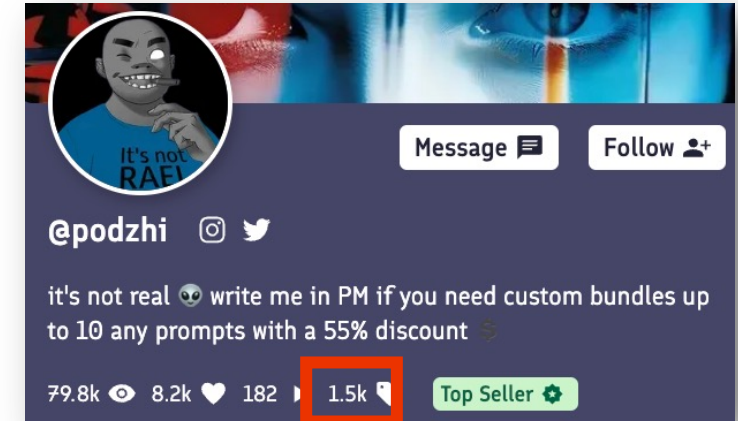
172.3k 12.8k 201 **2.8k** Top Seller



@mrdan Message Follow

Is artificial intelligence less than our intelligence?

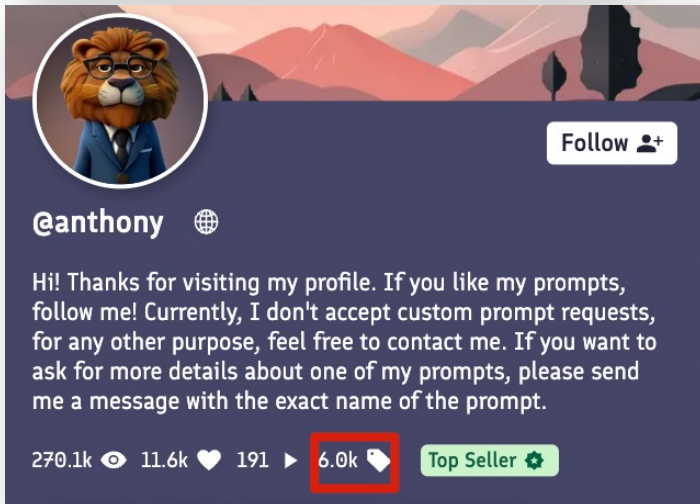
118.4k 5.5k 84 **2.4k** Top Seller



@podzhi Message Follow

it's not real write me in PM if you need custom bundles up to 10 any prompts with a 55% discount

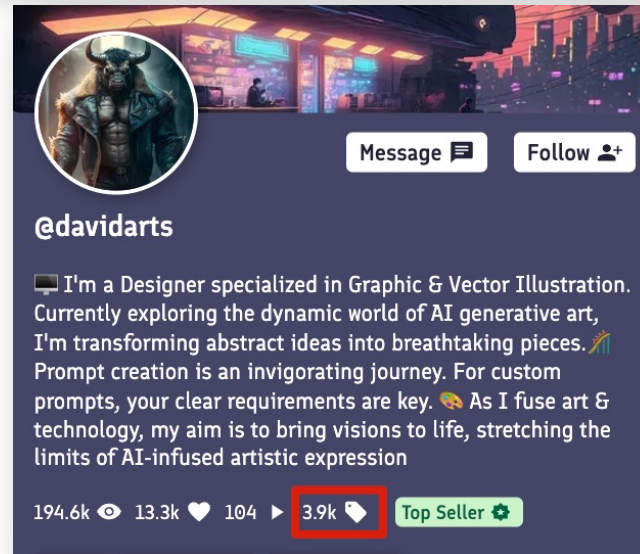
79.8k 8.2k 182 **1.5k** Top Seller



@anthony Follow

Hi! Thanks for visiting my profile. If you like my prompts, follow me! Currently, I don't accept custom prompt requests, for any other purpose, feel free to contact me. If you want to ask for more details about one of my prompts, please send me a message with the exact name of the prompt.

270.1k 11.6k 191 **6.0k** Top Seller



@davidarts Message Follow

I'm a Designer specialized in Graphic & Vector Illustration. Currently exploring the dynamic world of AI generative art, I'm transforming abstract ideas into breathtaking pieces. Prompt creation is an invigorating journey. For custom prompts, your clear requirements are key. As I fuse art & technology, my aim is to bring visions to life, stretching the limits of AI-infused artistic expression

194.6k 13.3k 104 **3.9k** Top Seller

- **Top 50** prompt engineers
- Sold **45,000 prompts** in 9 months
- Gained **~\$186,525** (estimated in 2023.10)



As Well As Prompt Trading Marketplaces

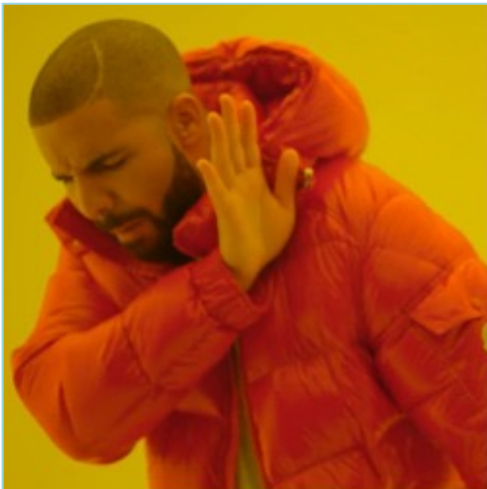
High-quality prompts become new commodities and are traded in new marketplaces

The image is a collage of several overlapping screenshots from different websites:

- PromptBase:** A dark-themed website with a search bar and navigation links for Models, Art, Logos, Graphics, Productivity, Marketing, Photography, and Games.
- PromptBase (Category View):** A screenshot showing a category of 'Logos' with a grid of various logo designs.
- PromptBase (Product Page):** A screenshot of a product page for a 'Free Marketplace' logo, featuring a 'BUY' button and a 'SUBSCRIBE' button.
- PromptBase (Product Page):** A screenshot of a product page for a 'Surrealistic Cores Digital Art' prompt, priced at \$4.99.
- PromptBase (Product Page):** A screenshot of a product page for a 'Cartoonish Whimsical Worlds' prompt, priced at \$2.99.
- PromptBase (Product Page):** A screenshot of a product page for a 'Painterly Street Visions' prompt, priced at \$4.99.
- PromptBase (Product Page):** A screenshot of a product page for a 'Messy Neon Art Prints' prompt, priced at \$3.99.
- PromptBase (Product Page):** A screenshot of a product page for a 'Brushstroke Soft Tones' prompt, priced at \$3.99.
- ChatX:** A dark-themed website with a search bar and a large graphic of two faces in profile, one purple and one orange, with the text 'AI' and 'Diffusion' overlaid.
- ChatX (Product Page):** A screenshot of a product page for a 'ChatGPT' prompt, priced at \$0.99.
- ChatX (Product Page):** A screenshot of a product page for a 'Stable Diffusion' prompt, priced at \$0.99.
- Promptrr.io:** A website with a dark theme and a blue/purple gradient background. It features a 'SUBSCRIBE' button and a 'BROWSE PROJECTS' button. The main text reads 'Convert future yield for early tokens'. Below this, it says 'Legato is a decentralized liquidity bootstrap protocol that enables the use of future staking yield to acquire early project tokens'. There are also buttons for 'Find a prompt' and 'Sell a prompt'.
- Promptrr.io (Product Page):** A screenshot of a product page for a 'Midjourney Prompt for a Social Media Advertisement' prompt, priced at \$0.99.
- Promptrr.io (Product Page):** A screenshot of a product page for a 'Midjourney Prompt for a Social Media Advertisement' prompt, priced at \$0.99.
- Promptrr.io (Product Page):** A screenshot of a product page for a 'Midjourney Prompt for a Social Media Advertisement' prompt, priced at \$0.99.
- Promptrr.io (Product Page):** A screenshot of a product page for a 'Midjourney Prompt for a Social Media Advertisement' prompt, priced at \$0.99.



Attack ?



Buy
prompts from
the marketplace

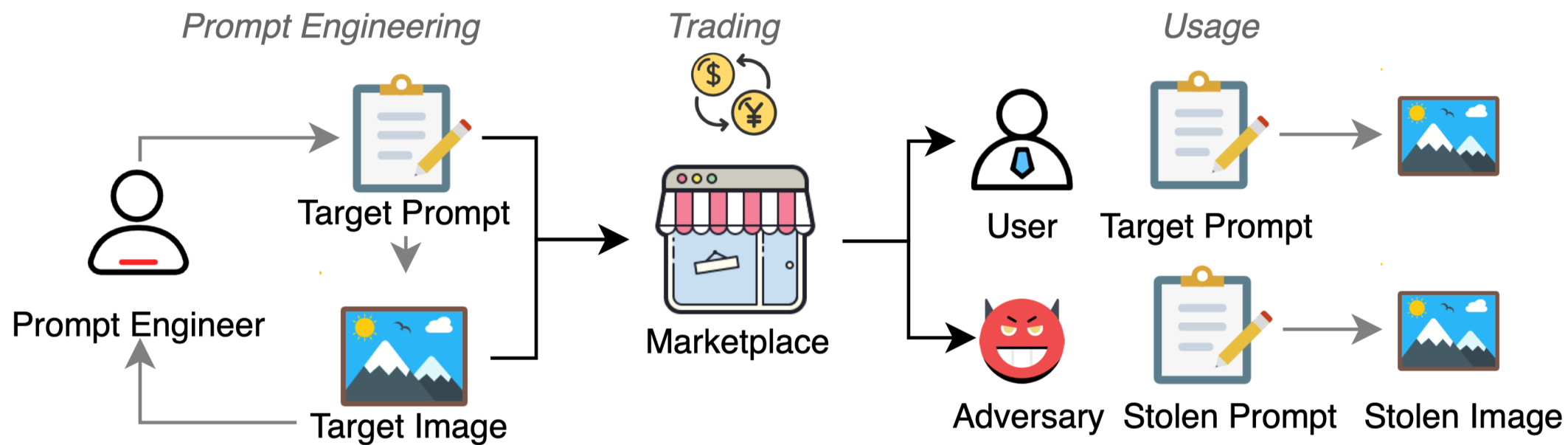


Steal
prompts
without paying

- Given an image generated by a text-to-image generation model, whether an attacker can infer its corresponding prompt?
- We name this novel attack as ***prompt stealing attack***



Threat Scenario

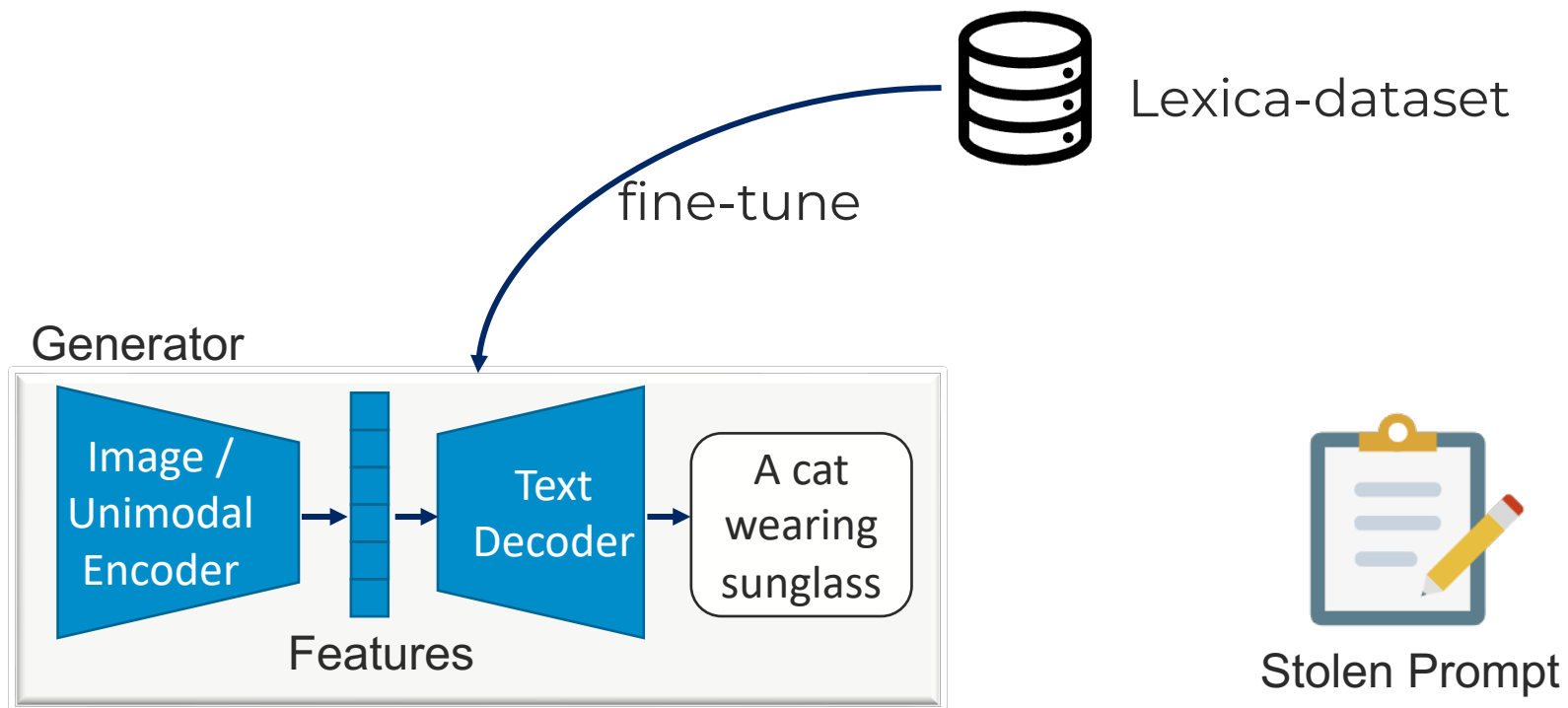




Attack Methodology



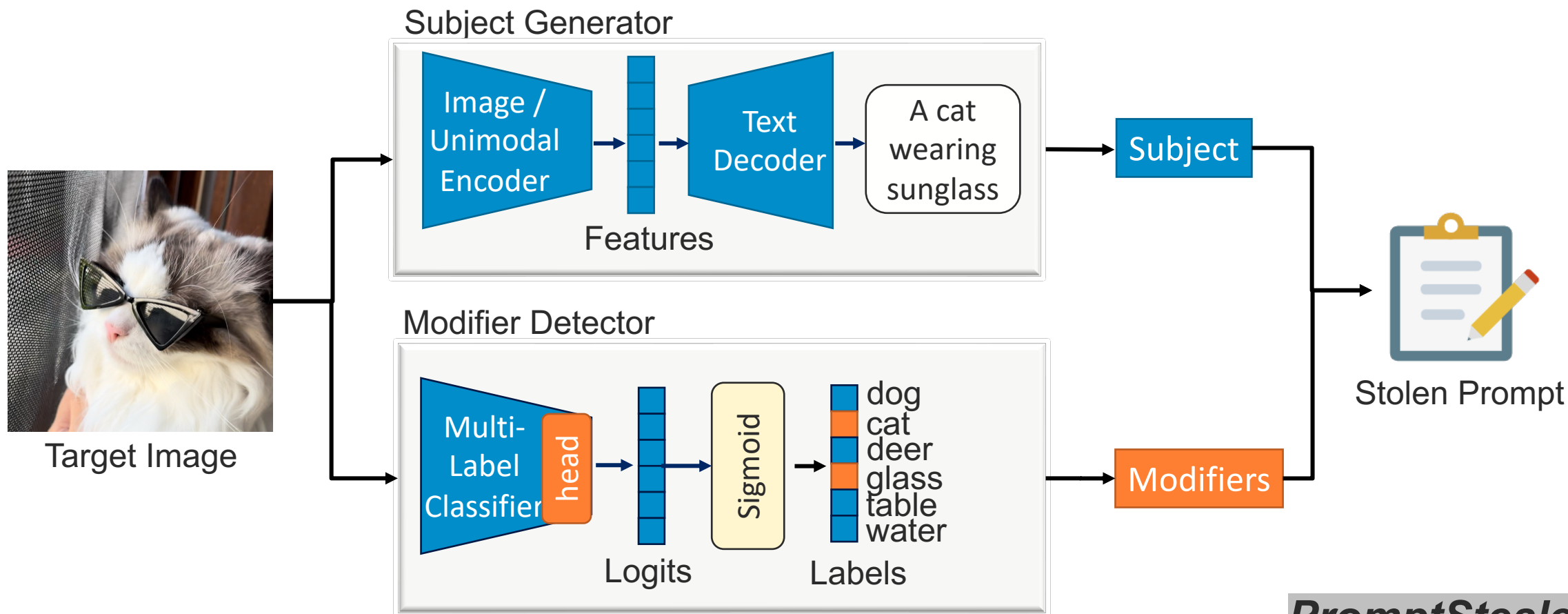
Target Image





Attack Methodology

- As our previous measurement shown that subject and modifiers are both important, thus...



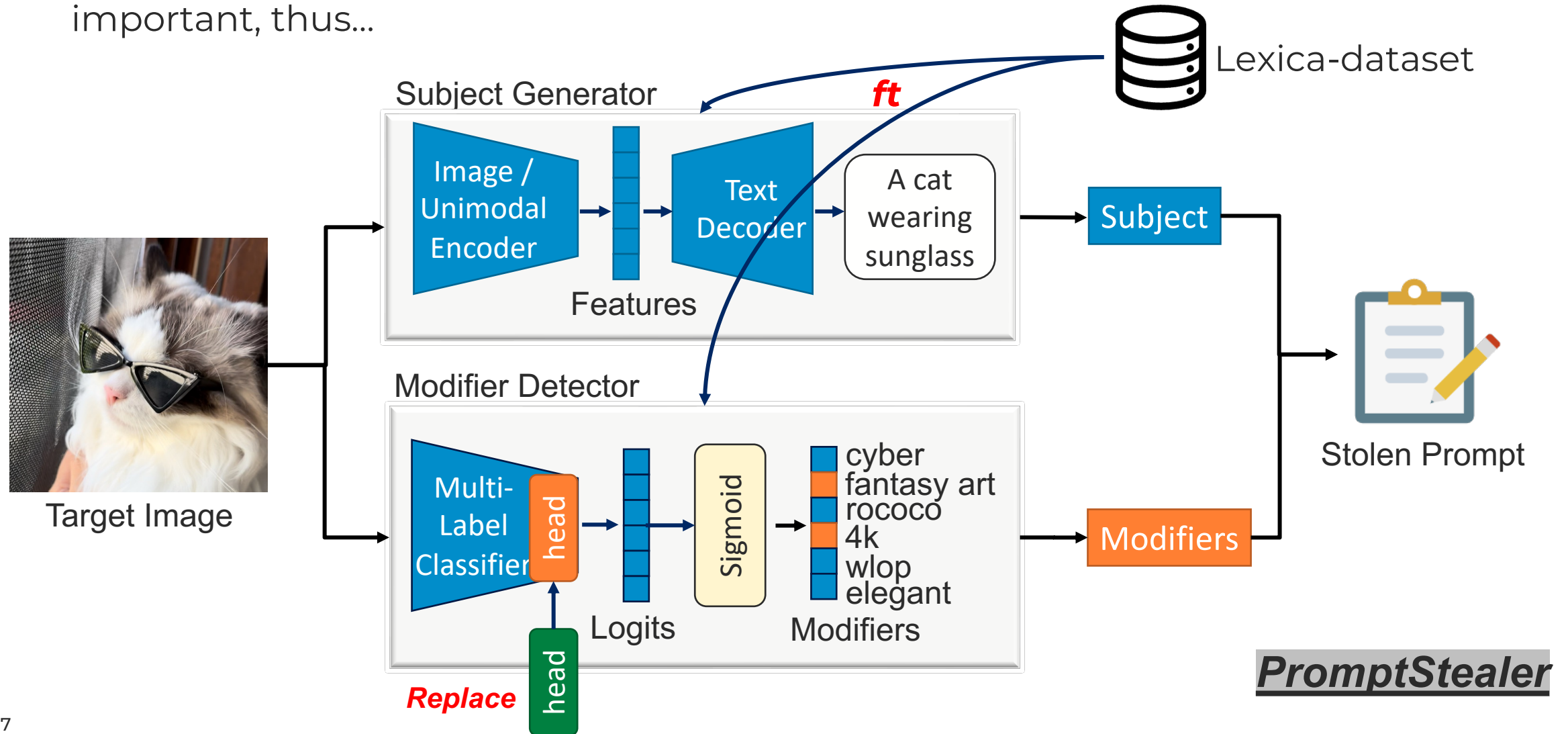
These are not modifiers

PromptStealer




Attack Methodology

- As our previous measurement shown that subject and modifiers are both important, thus...





Experimental Setup

- **Text-to-Image Generation Model:** Stable Diffusion 
- **Baseline:**
 - An image captioning model (blip)
 - Fine-tuned BLIP on Lexica-Dataset (our collected dataset)
 - CLIP Interrogator (an open-source tool)

- **Evaluation Metric**

- **Quantitative:**

- Semantic similarity
 - Modifier similarity
 - Image similarity
 - Pixel similarity



- **Qualitative:**

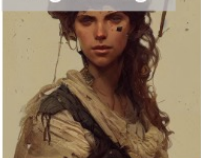
- Human-Rated Similarity





Qualitative Evaluation

Target Image



A full portrait of a beautiful post apocalyptic Bedouin explorer, intricate, elegant, highly detailed, digital painting, artstation, concept art, smooth, sharp focus, illustration, art by Krenz Cushart and Artem Demura and alphonse mucha

Image Captioning



a woman in a costume with a gun

Image Captioning (FT)



portrait of a post apocalyptic offworld adventurer, intricate, elegant, highly detailed, digital painting

CLIP Interrogator



a woman in a costume with a gun, a character portrait, jaimie jones, cgsociety, half the painting is glitched, woman in tattered clothes revealing body, female merchant, looks like alison brie, barbarian girl, stylized portrait

PromptStealer



a full portrait of a post apocalyptic offworld adventurer, artstation, highly detailed, concept art, sharp focus, digital painting, intricate, illustration, smooth, elegant, by krenz cushart and artem demura and alphonse mucha

Target Image



a study of cell shaded cartoon of the interior of a bioshock style art deco city, illustration, post grunge, concept art by josan gonzales and wlop, by james jean, victo ngai, david rubin, mike mignola, laurie greasley, highly detailed, sharp focus, trending on artstation, hq, deviantart, art by artgem

Image Captioning



a painting of a city at night

Image Captioning (FT)



a highly detailed matte painting of a steampunk cityscape by simon stalenhag

CLIP Interrogator



a painting of a city at night, cyberpunk art, stephan martiniere, cgsociety, anton fadeev and moebius, sketchfab, retro sci - fi : : a storyboard drawing, wlop : :

PromptStealer



a highly detailed illustration of a steampunk city, highly detailed, sharp focus, illustration, deviantart, by james jean, vibrant colors, by victo ngai, concept, wide shot, hq, laurie greasley, artgem, by mike mignola, by josan gonzales and wlop, david rubin



Quantitative Evaluation

- PromptStealer outperforms baseline both quantitatively and qualitatively

Method	Semantic	Modifier	Image	Pixel	Human
ImgCap	0.19	0.00	0.65	0.89	1.65
ImgCap (FT)	0.45	0.14	0.74	0.89	3.20
CLIP Interrogator	0.52	0.01	0.77	0.89	2.95
PromptStealer	0.70	0.43	0.80	0.90	4.45



Prompt Stealing Attacks on Other Model



Midjourney



DALL-E 2

Target Image



ultra realistic delorean dmc 5 with pop - up headlights drifting on ancient highway wreckage in space, dark cinematic, volumetric, realistic, 3 d render, realistic render, cinematic lighting, volumetric lighting, atmospheric, cinematic, unreal engine 5, unreal engine render, octane render, hd, photorealism, hyper realistic, 8 k

Stolen Images



a **delorean dmc** driving through a foggy field at night, **8k, octane render, smooth, cinematic lighting, cgsociety, volumetric lighting, hyper detailed, by craig mullins, by makoto shinkai, photorealism**

Target Image



ultra realistic delorean dmc 5 with pop - up headlights drifting on ancient highway wreckage in space, dark cinematic, volumetric, realistic, 3 d render, realistic render, cinematic lighting, volumetric lighting, atmospheric, cinematic, unreal engine 5, unreal engine render, octane render, hd, photorealism, hyper realistic, 8 k

Stolen Images



a delorean driving on a desert road, concept art, **8k, octane render, 4k, cinematic, cinematic lighting, hd, dramatic lighting, unreal engine 5, 3d, hyperrealistic, hyper realistic, photorealism, calm, unreal engine render**

Target Image



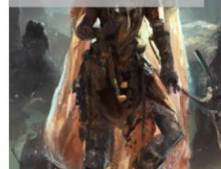
A full portrait of a beautiful post apocalyptic Bedouin explorer, intricate, elegant, highly detailed, digital painting, artstation, concept art, smooth, sharp focus, illustration, art by Krenz Cushart and Artem Demura and alphonse mucha

Stolen Images



a **beautiful post apocalyptic** woman sitting on a rock in the desert, **artstation, sharp focus, by greg rutkowski, cinematic, digital art, by wlop, hd, 3d, by craig mullins, pascal blanche**

Target Image



A full portrait of a beautiful post apocalyptic Bedouin explorer, intricate, elegant, highly detailed, digital painting, artstation, concept art, smooth, sharp focus, illustration, art by Krenz Cushart and Artem Demura and alphonse mucha

Stolen Images

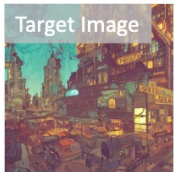


a **full body portrait of a beautiful post apocalyptic** offworld adventurer by greg rut, **artstation, concept art, 8k, sharp focus, digital art, high quality, oil painting, jama jurabaev, brush hard**



Compared PromptStealer to ChatGPT Plugin

- A popular prompt generator in the GPT store
- PromptStealer outperforms the GPT in recognizing the art style of the target image



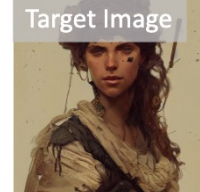
a study of cell shaded cartoon of the interior of a bioshock style art deco city, illustration, post grunge, concept art by josan gonzales and wlop, by james jean, victo ngai, david rubin, mike mignola, laurie greasley, highly detailed, sharp focus, trending on artstation, hq, deviantart, art by artgem



a highly detailed illustration of a steampunk city, highly detailed, sharp focus, illustration, deviantart, by james jean, vibrant colors, by victo ngai, concept, wide shot, hq, laurie greasley, artgem, by mike mignola, by josan gonzales and wlop, david rubin



Futuristic cityscape, warm pastel tones, bustling street scene, overcrowded architecture, flying vehicles, towering skyscrapers, neon signs, intricate details, sense of depth, urban congestion, clear sky, technological advancements, chaotic organization, blend of retro and future elements, sunlight casting long shadows, energy and dynamism, multiple layers of activity, diverse building designs, hints of vegetation, absence of people, mechanical complexity, digital billboards, cultural fusion, elevated roadways



A full portrait of a beautiful post apocalyptic Bedouin explorer, intricate, elegant, highly detailed, digital painting, artstation, concept art, smooth, sharp focus, illustration, art by Krenz Cushart and Artem Demura and alphonse mucha



a full portrait of a post apocalyptic offworld adventurer, artstation, highly detailed, concept art, sharp focus, digital painting, intricate, illustration, smooth, elegant, by krenz cushart and artem demura and alphonse mucha

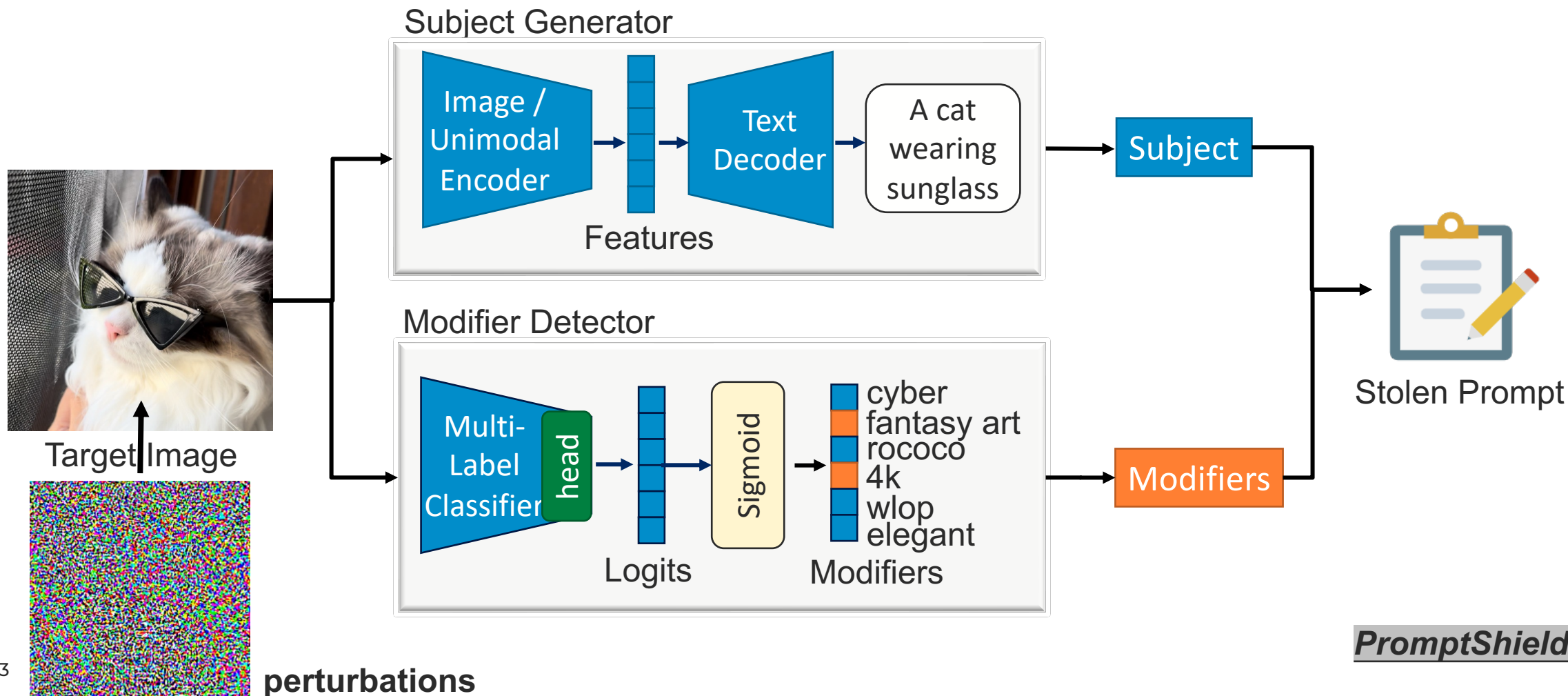


Determined gaze, survivalist attire, smudge of war paint, tousled brown hair, futuristic headgear, hint of a scar, neutral tones, canvas of resilience, makeshift bandana, look of resolve, weathered clothing, ambient cream background, rugged beauty, adorned with gadgets, post-apocalyptic vibe, warrior stance, young but hardened, resourceful character, soft lighting, story in her eyes, minimalist portrait, earthy palette, essence of fortitude, lone fighter, subtle defiance, character of depth



Defense

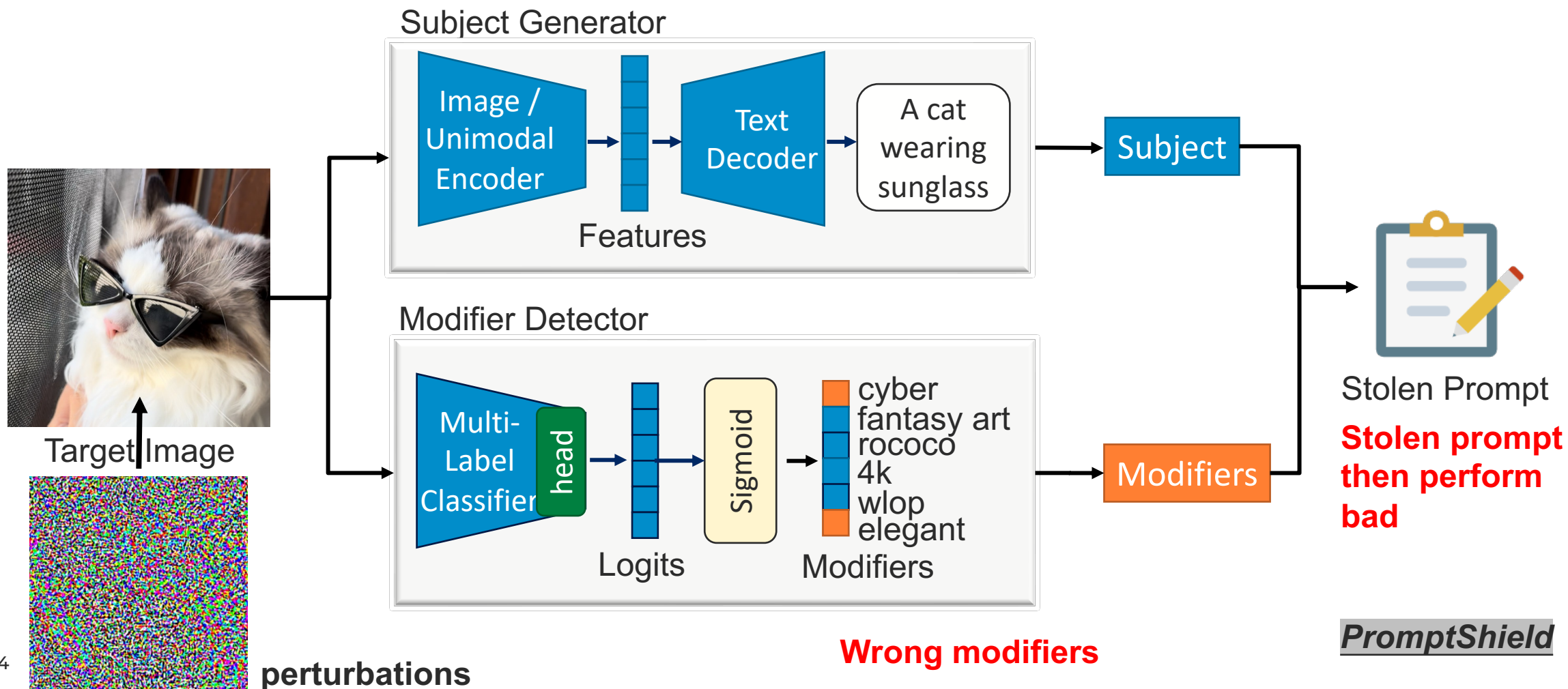
- Can we reduce PromptStealer's machine learning models' performance?





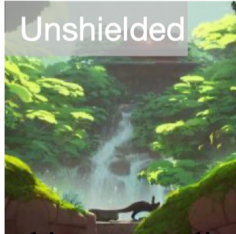
Defense

- Can we reduce PromptStealer's machine learning models' performance?





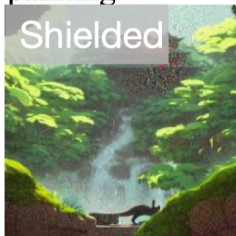
Qualitative Evaluation



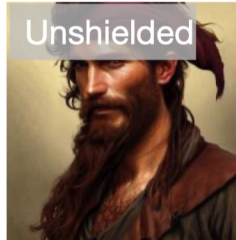
a big cat guarding the entrance to a temple, lush vegetation, waterfalls, cory loftis, james gilleard, atey ghailan, makoto shinkai, goro fujita, character art, rim light, exquisite lighting, clear focus, very coherent, plain background, soft painting



a highly detailed matte painting of a japanese temple in a lush forest by studio ghibli, by makoto shinkai, by studio ghibli, rim light, by james gilleard, very coherent, by atey ghailan, plain background, clear focus, by goro fujita, exquisite lighting, cory loftis, soft painting, lush vegetation



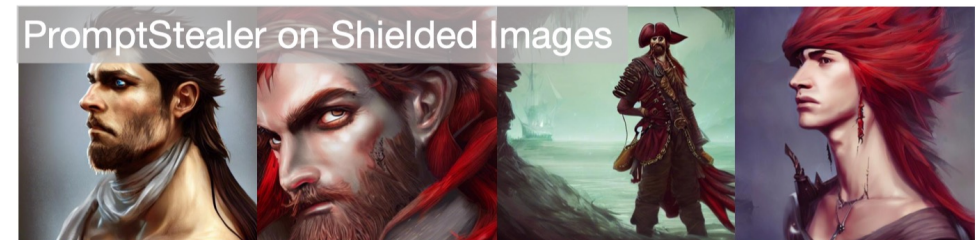
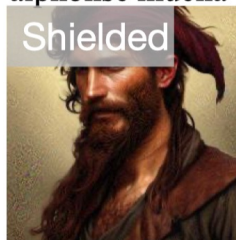
a cat in a japanese garden, rim light, very coherent, plain background, character art, clear focus, exquisite lighting, cory loftis, soft painting, lush vegetation, waterfalls



portrait of a young ruggedly handsome but cantankerous pirate, male, masculine, upper body, red hair, long hair, d & d, fantasy, bashful smirk, intricate, elegant, highly detailed, digital painting, artstation, concept art, matte, sharp focus, illustration, art by artgerm and greg rutkowski and alphonse mucha



portrait of a rugged pirate, artstation, highly detailed, concept art, sharp focus, digital painting, intricate, illustration, elegant, fantasy, by artgerm and greg rutkowski and alphonse mucha, matte, d & d, male, long hair, upper body, masculine, red hair



a pirate, artstation, highly detailed, concept art, sharp focus, digital painting, intricate, illustration, elegant, fantasy, matte, d & d, male, long hair, upper body, masculine, red hair



Conclusion

- The first study on prompt stealing attack
- This work has also been recognized in ***Microsoft Vulnerability Severity Classification for AI Systems¹***
- PromptStealer outperforms baseline methods both quantitatively and qualitatively
- We also make the first attempt to mitigate prompt stealing attacks by proposing *PromptShield*



<https://github.com/verazuo/prompt-stealing-attack>



https://huggingface.co/datasets/vera365/lexica_dataset



xinyueshen.me

^[1] <https://www.microsoft.com/en-us/msrc/aibugbar>