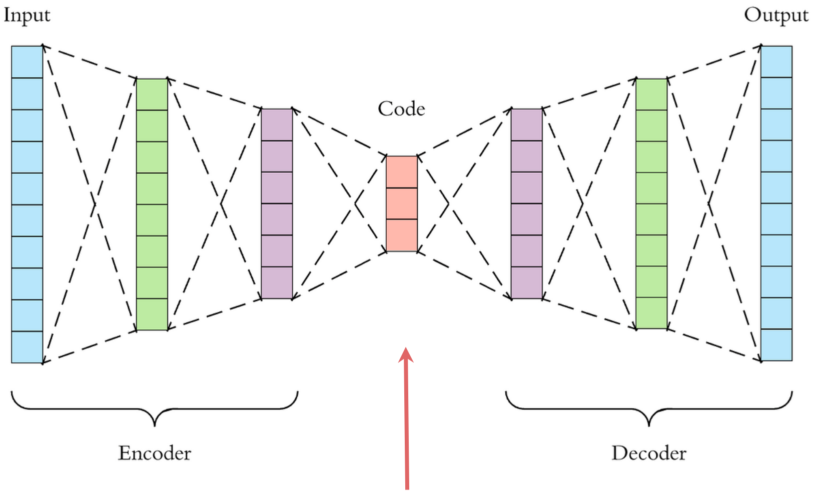
Motivation

Deep Embeddings

Representing an image by continuous embeddings from deep networks.



visual representation

Autoencoder, 2007

iBOT, 2021

MaskFeat, 2022

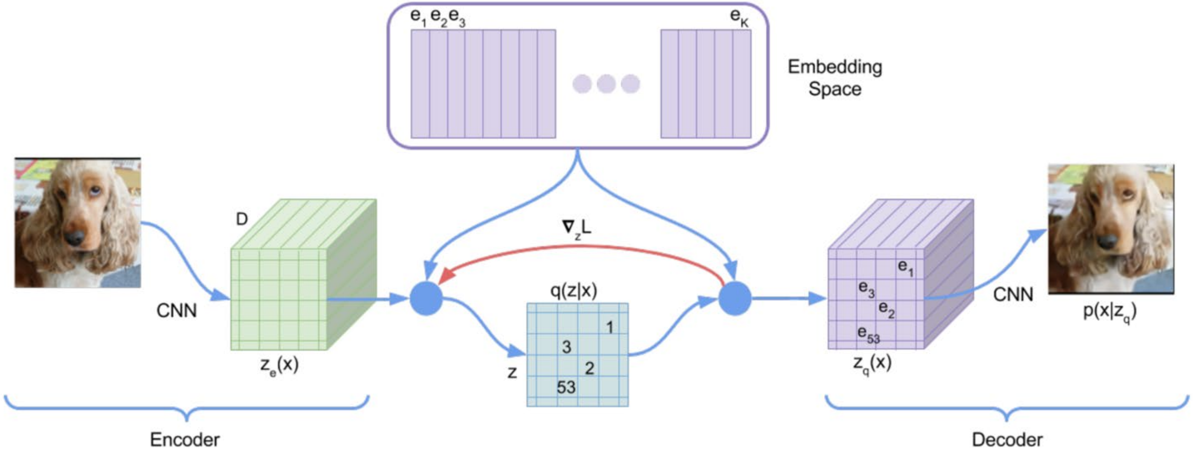
DiffMAE, 2023

……

Motivation

Deep Embeddings vs. Quantized Tokens

Representing an image by quantized tokens from deep networks.



visual representation

VQ-VAE, 2017

DALL-E, 2021

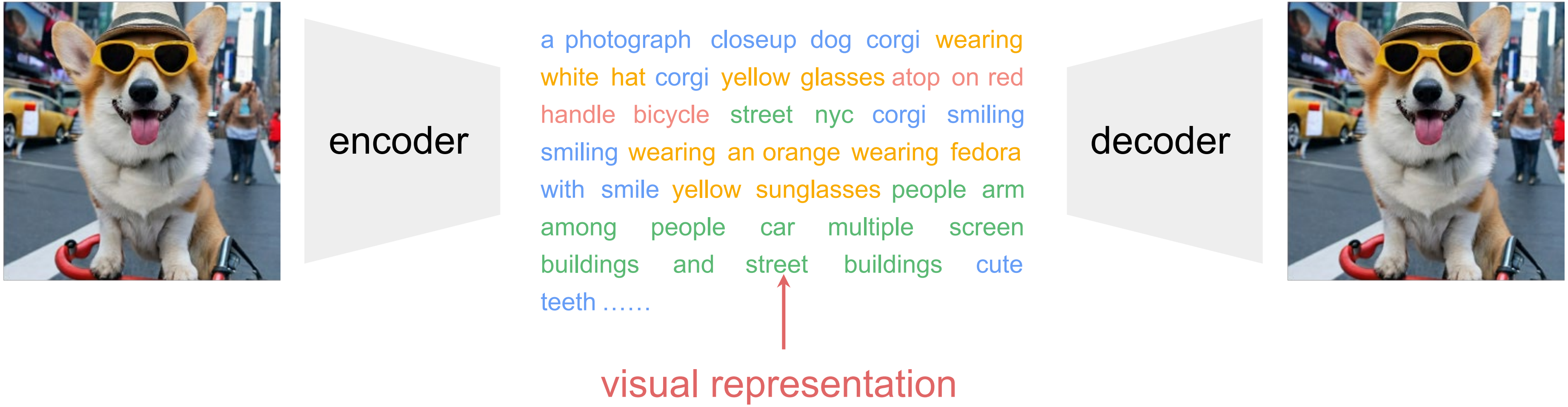
Parti, 2022

……

Motivation

Deep Embeddings vs. Quantized Tokens vs. Text

Can we represent an image as text?



encoder

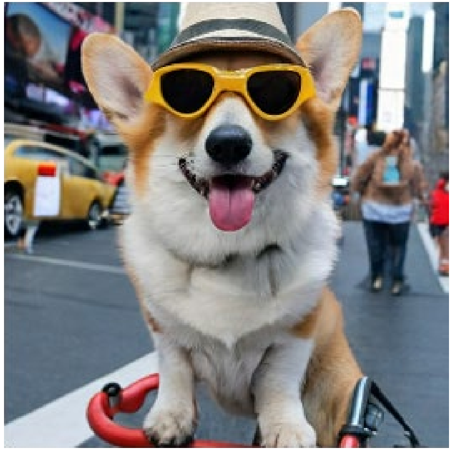
a photograph closeup dog corgi wearing white hat corgi yellow glasses atop on red handle bicycle street nyc corgi smiling smiling wearing an orange wearing fedora with smile yellow sunglasses people arm among people car multiple screen buildings and street buildings cute teeth ……

decoder

visual representation

Example

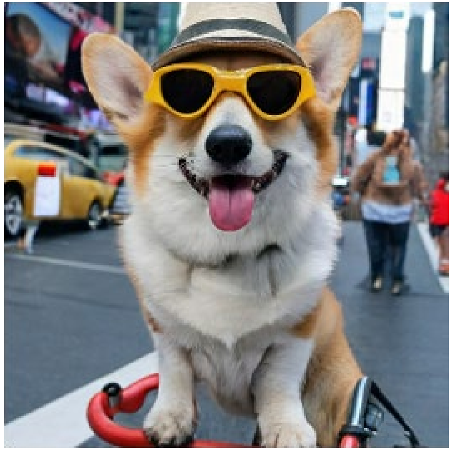
De-Diffusion Text



an attributing provided closeup dog corgi wearing wht yellow hat corgi and hat on glasses aboard an on red handle cart street nyc oscanadian corgi corgi presented description called an dog dog shown standing smiling incorporating wearing an wearing harnehandle fedora with smile yellow sunglasses yellow sunglasses and people and street wth foreground with people people among people cars multiple billboards cityscape and street billboards asphalt street cute screenshot street

Example

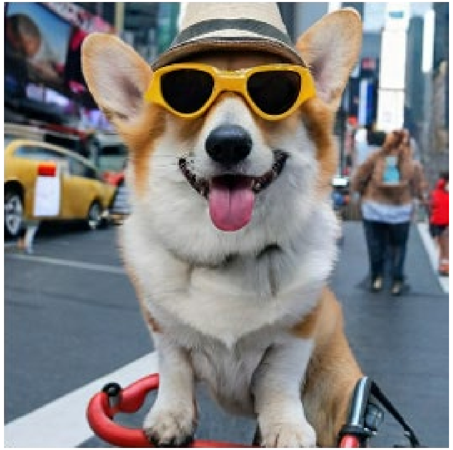
De-Diffusion Text



an attributing provided closeup dog corgi wearing wht yellow hat corgi and hat on glasses aboard an on red handle cart street nyc oscanadian corgi corgi presented description called an dog dog shown standing smiling incorporating wearing an wearing harnehandle fedora with smile yellow sunglasses yellow sunglasses and people and street wth foreground with people people among people cars multiple billboards cityscape and street billboards asphalt street cute screenshot street

Example

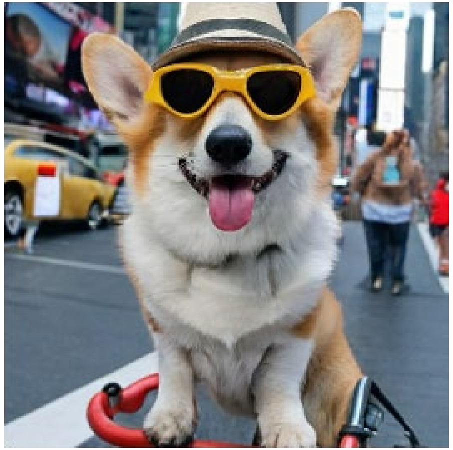
De-Diffusion Text



an attributing provided closeup dog corgi wearing wht yellow hat corgi and hat on glasses aboard an on red handle cart street nyc oscanadian corgi corgi presented description called an dog dog shown standing smiling incorporating wearing an wearing harnehandle fedora with smile yellow sunglasses yellow sunglasses and people and street wth foreground with people people among people cars multiple billboards cityscape and street billboards asphalt street cute screenshot street

Example

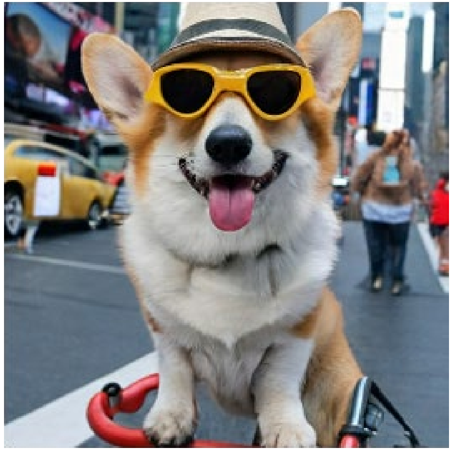
De-Diffusion Text



an attributing provided closeup dog corgi wearing wht yellow hat corgi and hat on glasses aboard an on red handle cart street nyc oscanadian corgi corgi presented description called an dog dog shown standing smiling incorporating wearing an wearing harnehandle fedora with smile yellow sunglasses yellow sunglasses and people and street wth foreground with people people among people cars multiple billboards cityscape and street billboards asphalt street cute screenshot street

Example

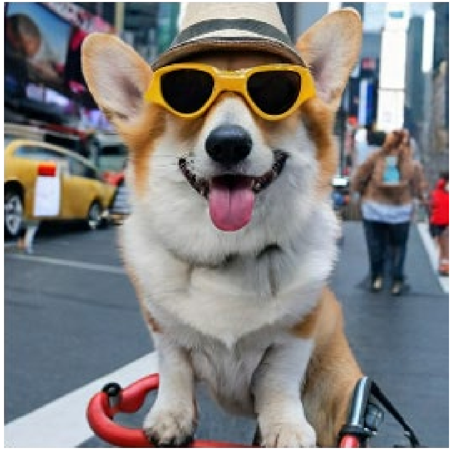
De-Diffusion Text



an attributing provided closeup dog corgi wearing wht yellow hat corgi and hat on glasses aboard an on red handle cart street nyc oscanadian corgi corgi presented description called an dog dog shown standing smiling incorporating wearing an wearing harnehandle fedora with smile yellow sunglasses yellow sunglasses and people and street wth foreground with people people among people cars multiple billboards cityscape and street billboards asphalt street cute screenshot street

Example

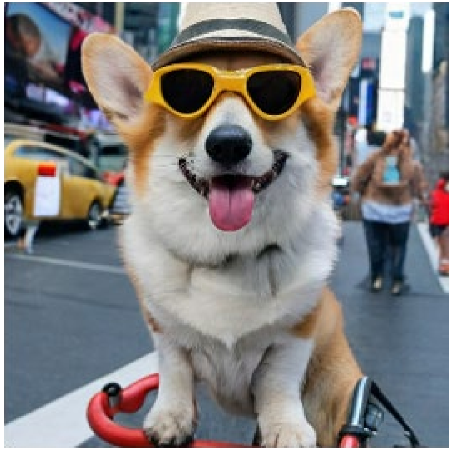
De-Diffusion Text



an attributing provided closeup dog corgi wearing wht yellow hat corgi and hat on glasses aboard an on red handle cart street nyc oscanadian corgi corgi presented description called an dog dog shown standing smiling incorporating wearing an wearing harnehandle fedora with smile yellow sunglasses yellow sunglasses and people and street wth foreground with people people among people cars multiple billboards cityscape and street billboards asphalt street cute screenshot street

Example

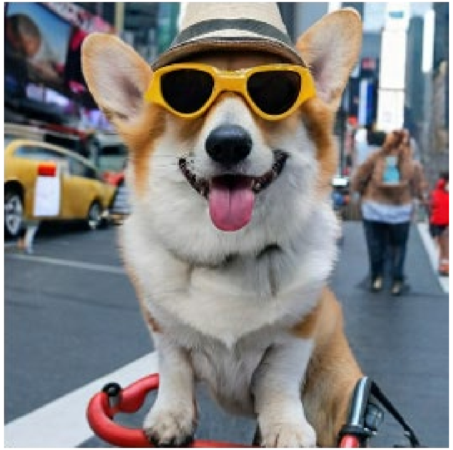
De-Diffusion Text



an attributing provided closeup dog corgi wearing wht yellow hat corgi and hat on glasses aboard an on red handle cart street nyc oscanadian corgi corgi presented description called an dog dog shown standing smiling incorporating wearing an wearing harnehandle fedora with smile yellow sunglasses yellow sunglasses and people and street wth foreground with people people among people cars multiple billboards cityscape and street billboards asphalt street cute screenshot street

Example

De-Diffusion Text

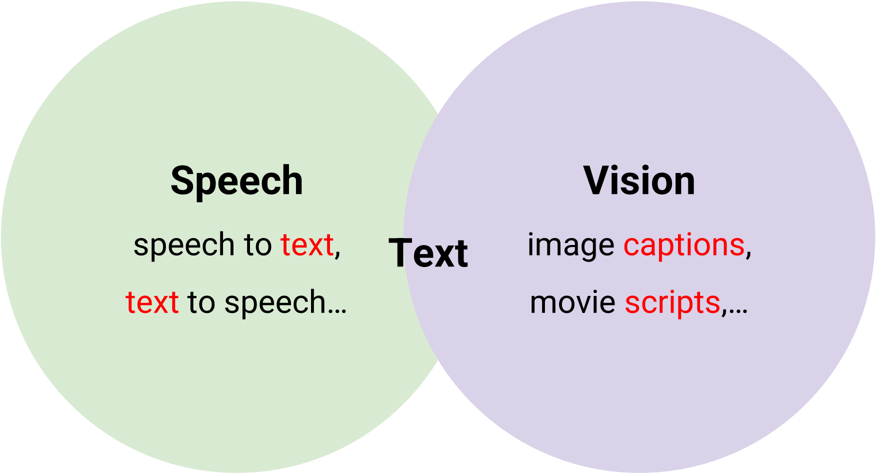


an attributing provided closeup dog corgi wearing wht yellow hat corgi and hat on glasses aboard an on red handle ca eet nyc oscanadian corgi corgi presented des g dog shown standing smiling an wearing harne ha ellow sunglasses yellow sunglasses and p eet wth foreground with people people among people cars multiple billboards cityscape and street billboards asphalt street cute screenshot street

Scrambled Caption

Motivation

Text as Cross -Modal Representation



Speech

speech to text,

text to speech…

Text

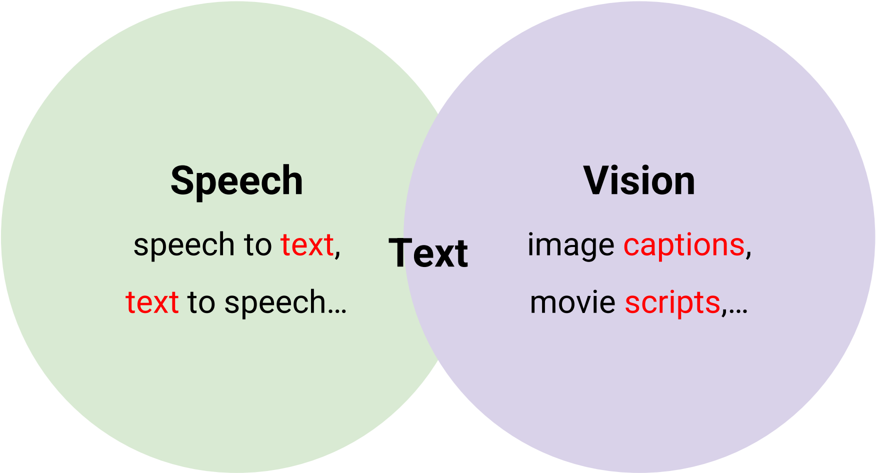
Vision

image captions,

movie scripts,…

Humans are used to encode many different modalities into text.

Text as Cross -Modal Representation



Speech

speech to text,

text to speech...

Text

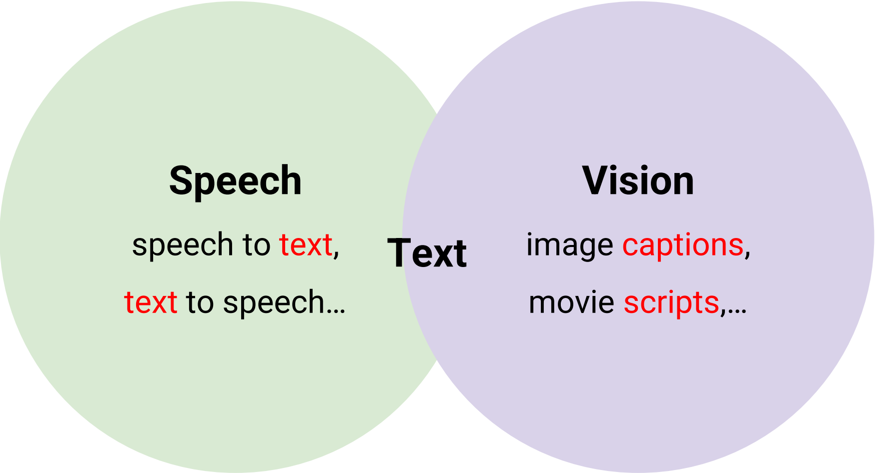
Vision

image captions,

movie scripts,...

Humans are used to encode many different modalities into text.

Text as Cross -Modal Representation



Speech

speech to text,

text to speech…

Text

Vision

image captions,

movie scripts,…

“In daily life, language often acts as an interface to higher -level cognition.”

Why not?

Text as Cross-Modal Representation

Why not captions?

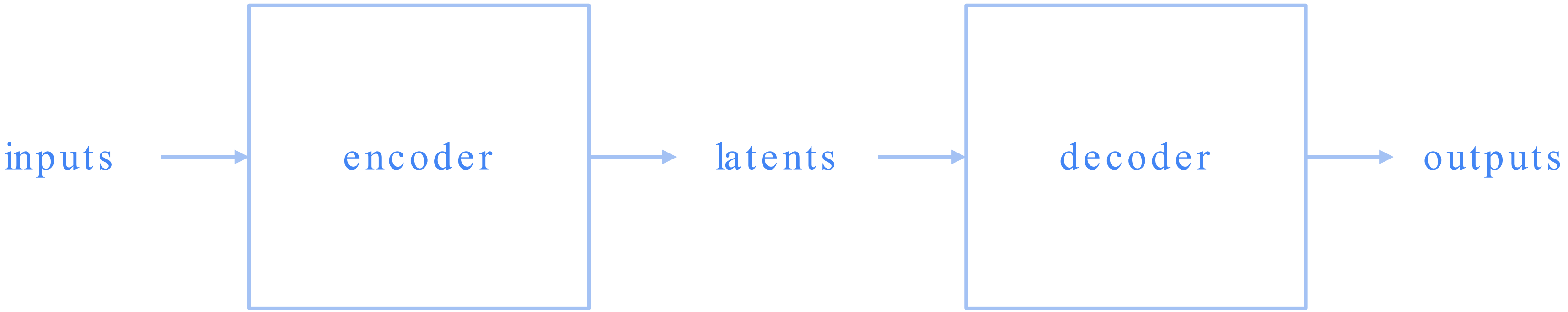
- Captions are usually not as comprehensive.

Why not deep embeddings?

- Interpretability
- Text can be directly used in LLMs &text-to-img models, no need of model tuning

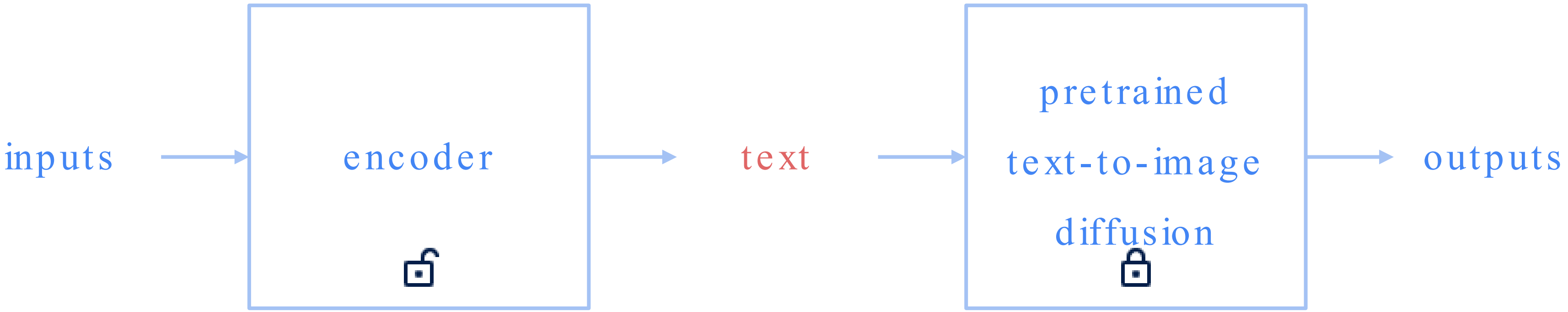
Method

An Autoencoder



- Training only required images.

Diffusion as Decoder

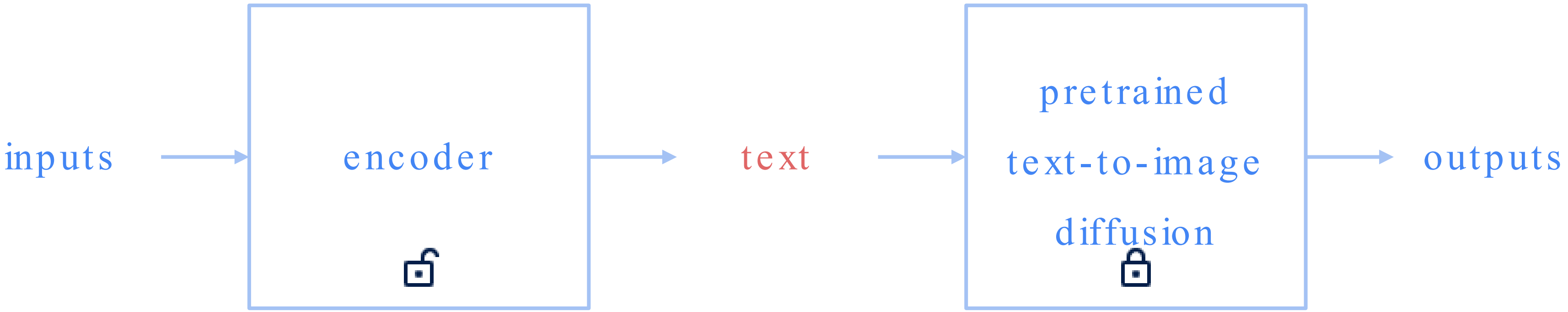


- Training only required images.
- Unpacking the knowledge encapsulated within the text -to-image generative models.

Method

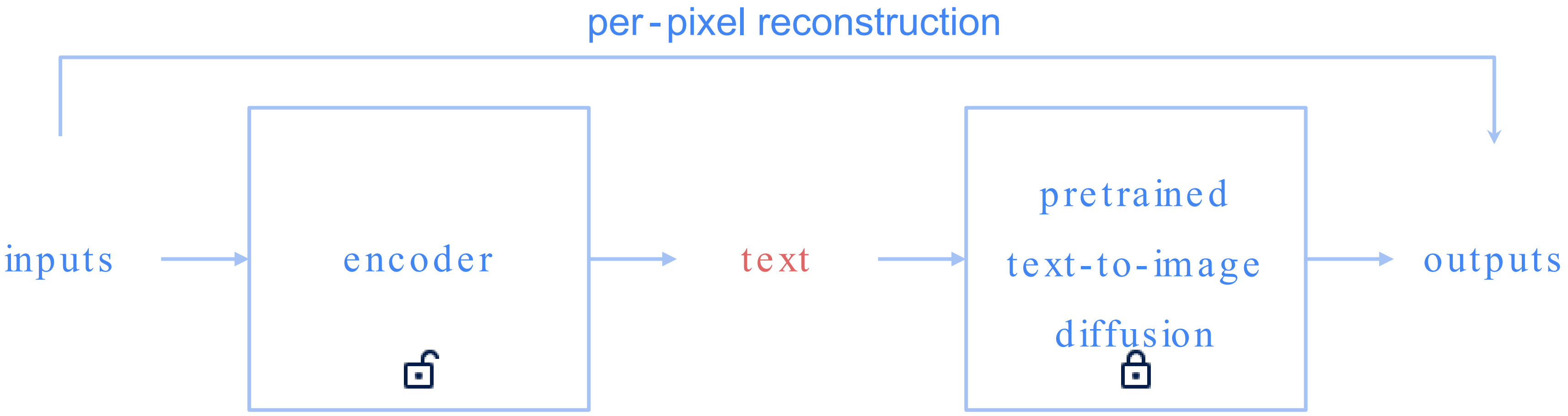
Diffusion as Decoder

Avocado chair


inputs → encoder → text → pretrained text-to-image diffusion → outputs

- Training only required images.
- Unpacking the knowledge encapsulated within the text -to-image generative models.

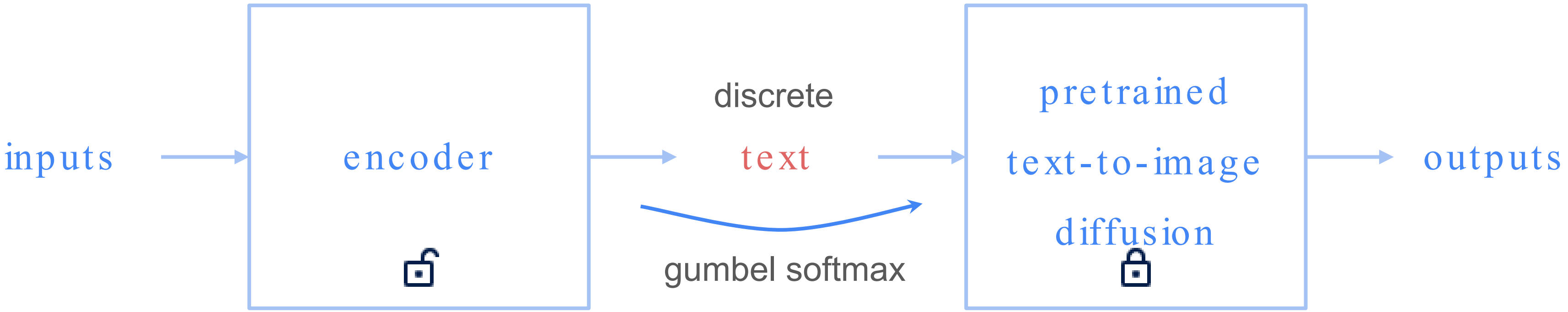
Diffusion as Decoder



- Training only required images.
- Unpacking the knowledge encapsulated within the text -to-image generative models.
    - To minimize the reconstruction error, the text has to be comprehensive .

Method

Full Picture

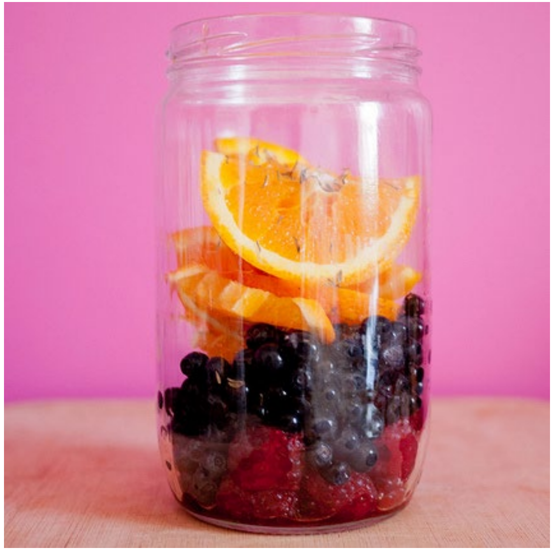


- Training only required images.
- Unpacking the knowledge encapsulated within the text -to-image generative models.
    - To minimize the reconstruction error, the text has to be accurate and comprehensive.
- Discrete text tokens, trained with gumbel softmax.

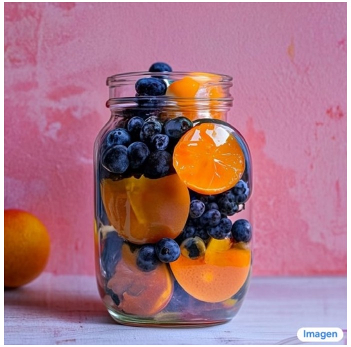
Application #1

Transferable Text-to-Image Prompt

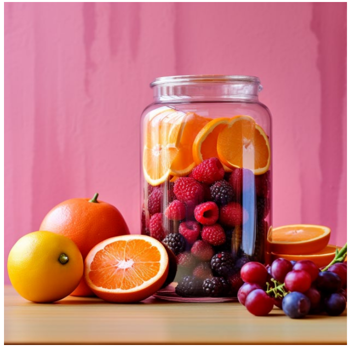
original



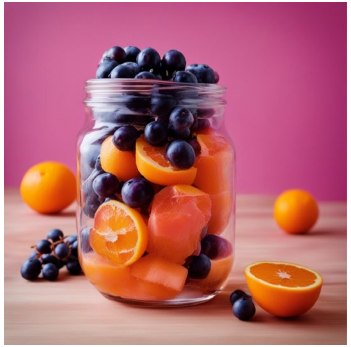
Imagen



Midjourney



Stable Diffusion XL

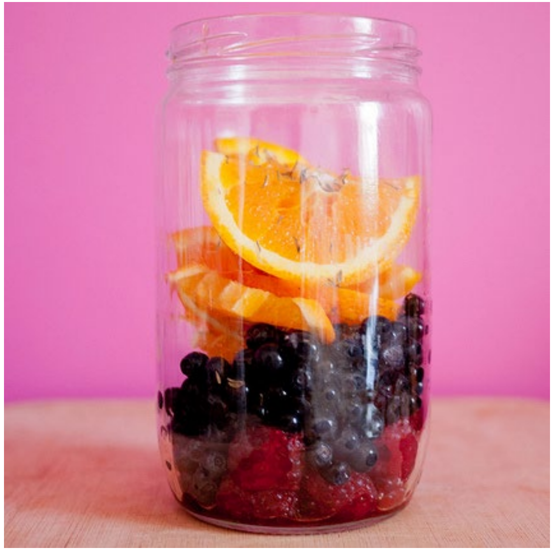


an davilishlish blog closeup berries jar through large refrerefre jar glass jar each other glass on an on peach hardwood closeup glass homemade mixed glass jar called relating called an oranges fruit shown slices each other containing relating an orange orange slices slices between black grapes open chunks orange oranges and berry black blueberry consist though towards pink closeup facing that background pink wall background pink pink wall wall closeup chia grapes recipe

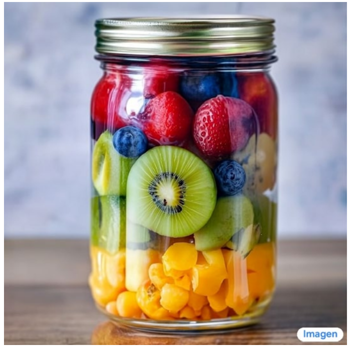
Application #1

Compare with Human Annotation

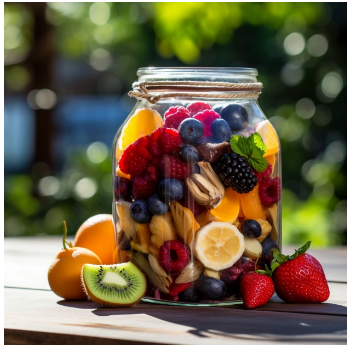
original



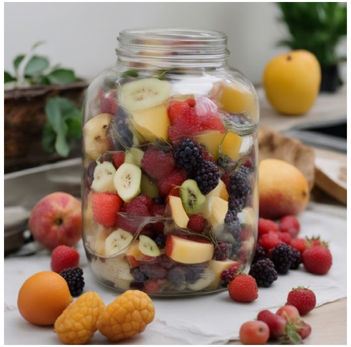
Imagen



Midjourney



Stable Diffusion XL



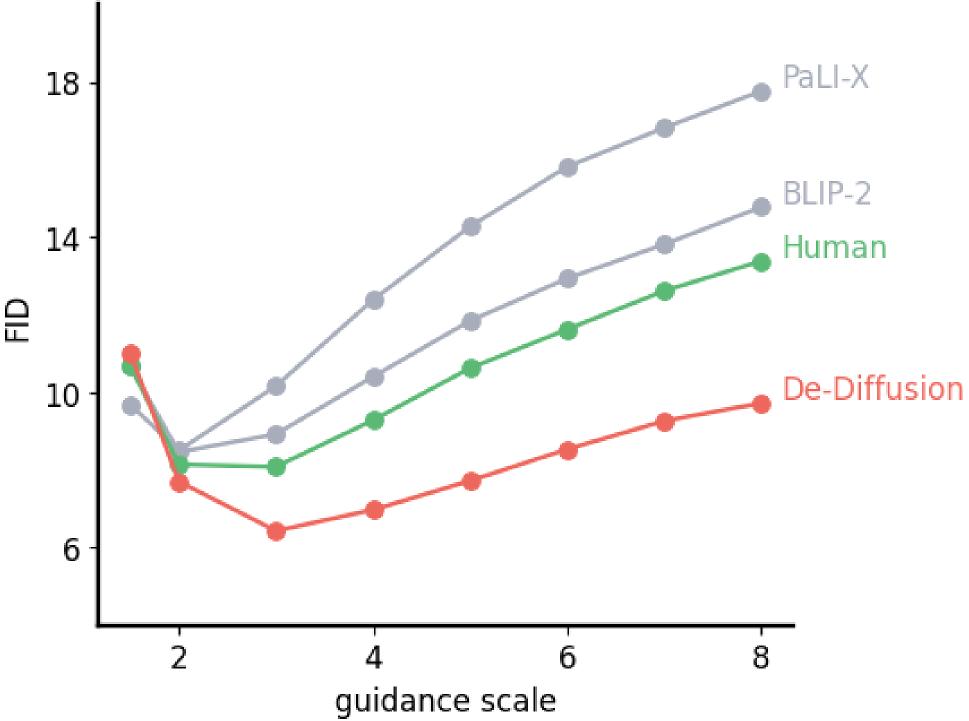
[Human-annotated caption] A jar filled with different types of fruit on a table.

Why is this different from captions?

- Captions are usually not as comprehensive.

Application #1

Benchmark on a Third -Party Diffusion Model



FID (↓): Similarity measure

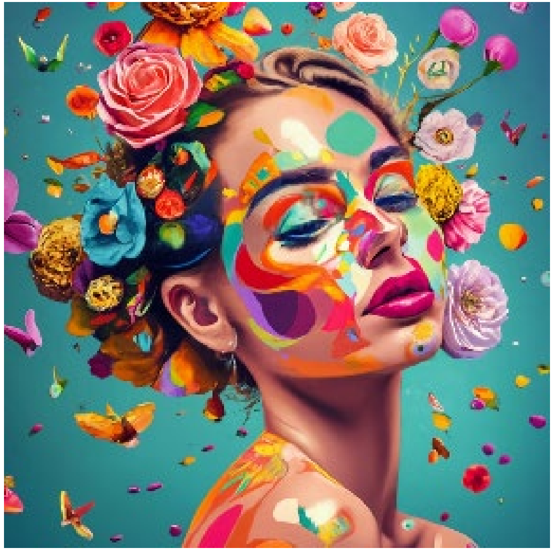
Diffusion Model: Stable Diffusion v2

Dataset: MS-COCO Val2014 30K

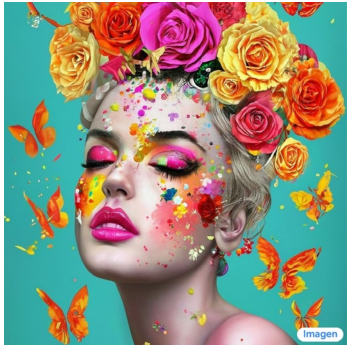
Application #1

Prompt for Synthetic Images

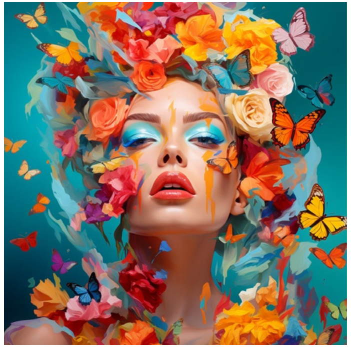
original



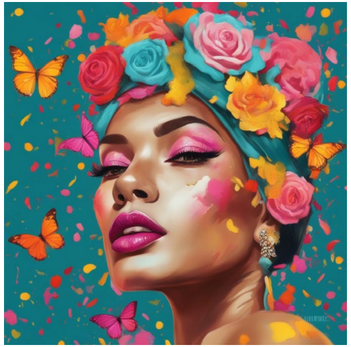
Imagen



Midjourney



Stable Diffusion XL

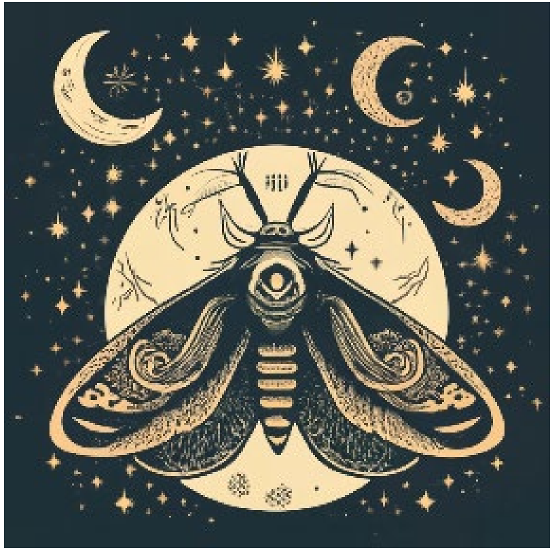


an art rha digitally s art illustration woman face wearing colorful colorful paints face painted head pink lipstick though an among colourful confetti confetti realism pin up os janu monroe monroe resembrelating called an face woman shown face smelling upwards multiple an colorful florals roses hats above many paints with earrings turmeric makeup brightly orange red pink wth scattered among yellow oranges flying flying butterflies teal background on teal blue background lips eyebrow hadid cg poster

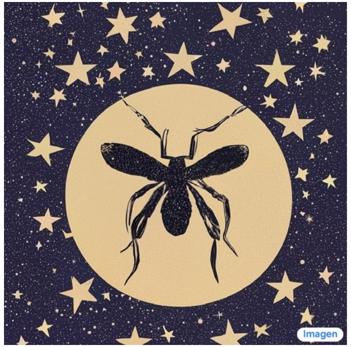
Application #1

Prompt for Synthetic Images

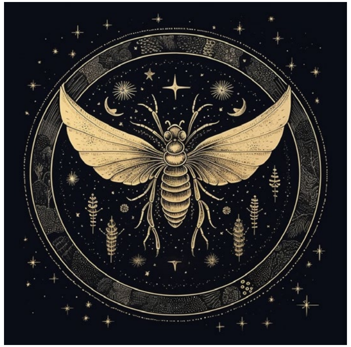
original



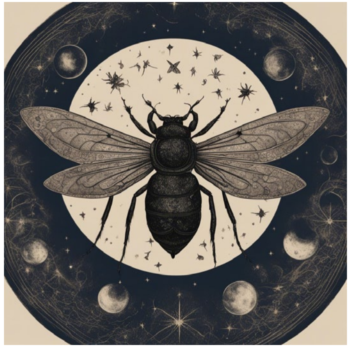
Imagen



Midjourney



Stable Diffusion XL

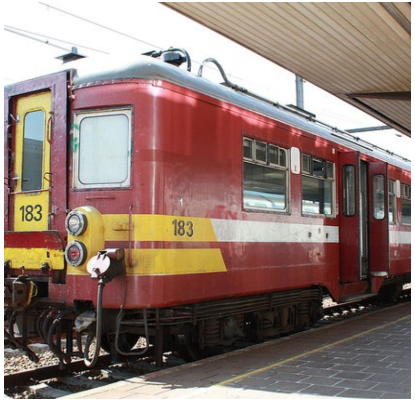


an illustration albu etching vscocam illustration intricate insect heavily black intricate intricate insect insect crest intricate crest on an behind lit circular moon intricate folk os intricate insect insect form a exhibiting called an intricate insect shown frontal frontal surrounded amongst an lit many crescent moons besides scattered stars and stars and moons past gold beige navy amongst beside among and crescent beside and crescent navy stars on dark navy background night stars bohemian etching logo

Application #2

Visual Question Answering

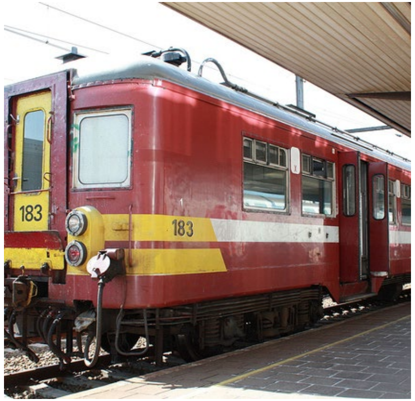
[Prompt] Answer the question given the context.



Question: What other big vehicle is often painted about the same shade as this vehicle?

Application #2

Visual Question Answering



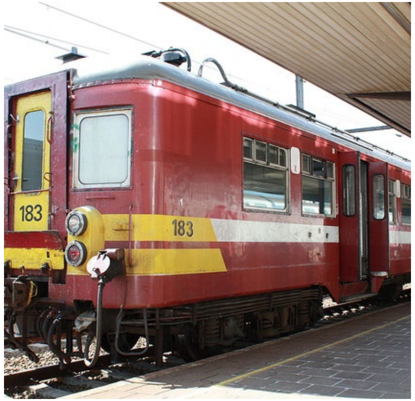
[Prompt] Answer the question given the context.

a colcandidenver lantic closeup former recent **train train** parked tradition enclosed **metrotram in a red livery** it on railroad platform containing **wearing a yellow pol** surround a knob beside platform near a under platform shelter right there and roof shadows and platform and tracks etc wore worn worn mau maroon brown **white stripes markings** content worn **yellow yellow stripes train** pretoria namibia railway platform train operator worn brown windows platform platform

Question: What other big vehicle is often painted about the same shade as this vehicle?

Application #2

Visual Question Answering with LLM



[Prompt] Answer the question given the context.

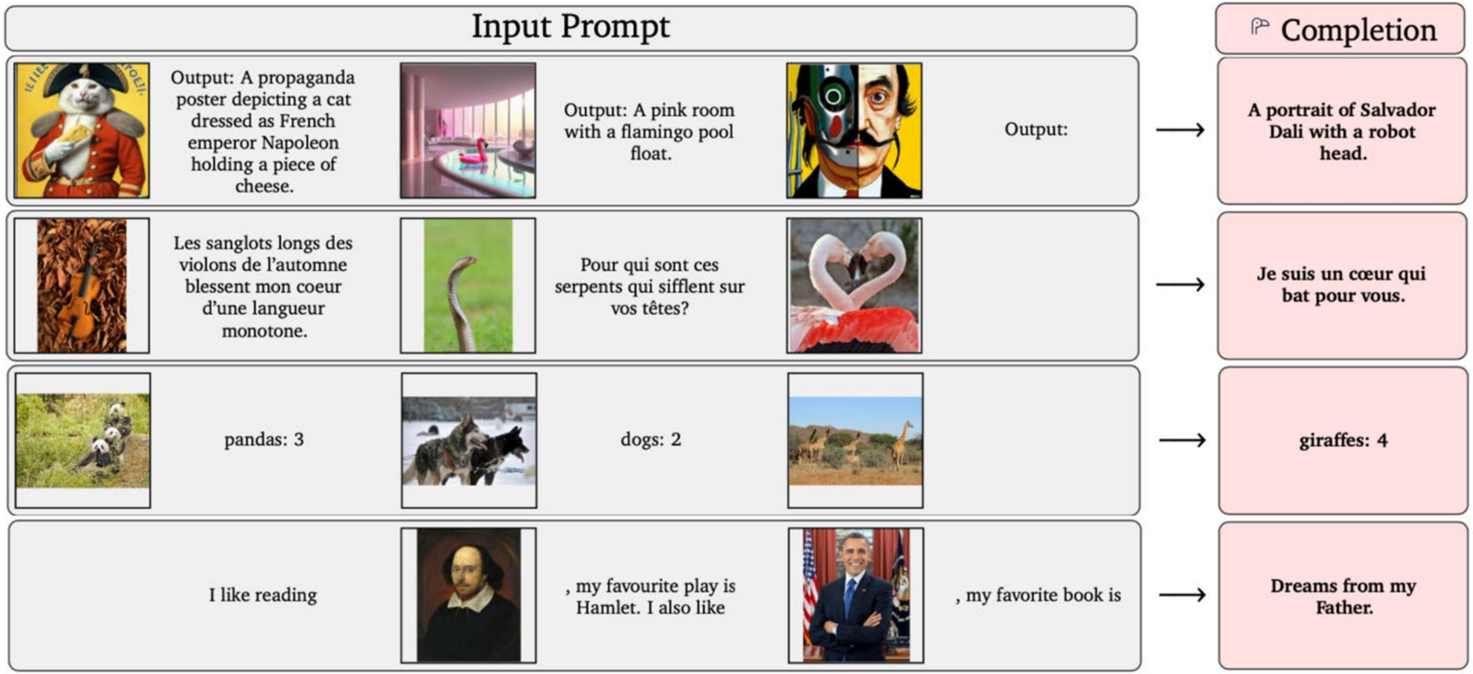
a colcandidenver lantic closeup former recent **train train** parked tradition enclosed **metrotram in a red livery** it on railroad platform containing **wearing a yellow pol** surround a knob beside platform near a under platform shelter right there and roof shadows and platform and tracks etc wore worn worn mau maroon brown **white stripes markings** content worn **yellow yellow stripes train** pretoria namibia railway platform train operator worn brown windows platform platform

Question: What other big vehicle is often painted about the same shade as this vehicle?

[LLM PaLM 2] Fire truck.

Application #2

Multi -Modal Few -Shot Learning



Input Prompt

Completion

Output: A propaganda poster depicting a cat dressed as French emperor Napoleon holding a piece of cheese.

Output: A pink room with a flamingo pool float.

Output: → A portrait of Salvador Dali with a robot head.

Les sanglots longs des violons de l'automne blessent mon coeur d'une langueur monotone.

Pour qui sont ces serpents qui sifflent sur vos têtes?

→ Je suis un cœur qui bat pour vous.

pandas: 3

dogs: 2

→ giraffes: 4

I like reading

, my favourite play is Hamlet. I also like

, my favorite book is → Dreams from my Father.

Flamingo, DeepMind, 2022

Application #2

Modular, Better

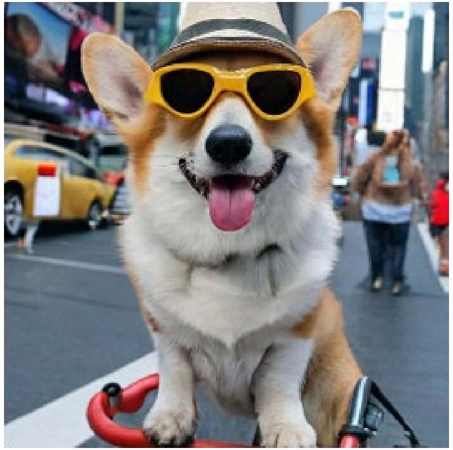
Model	VQAv2	OKVQA	COCO Caption
Flamingo-80B	56.3/67.6	50.6/57.8	84.3/113.8
De-Diffusion with PaLM 2	67.2/68.4	57.0/60.6	88.5/103.7

Representing images with text provides flexibility for using LLMs for multimodal tasks.

- easy exploitation of advanced LLMs: no joint training required, just inference;
- easy data collection: no interleaved image-text data required, just images.

Application #3

Interactions with ChatGPT



C

Given an image with keywords:

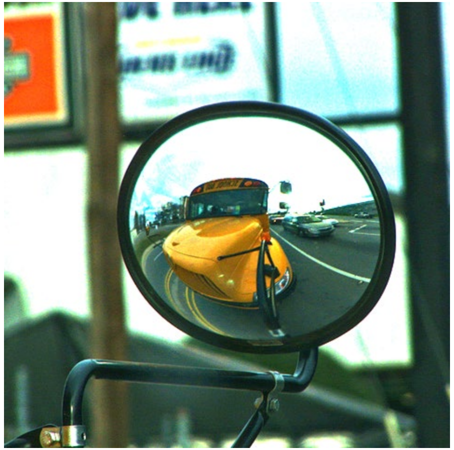
“an attributing provided closeup dog corgi wearing wht yellow hat corgi and hat on glasses aboard an on red handle cart street nyc oscanadian corgi corgi presented description called an dog dog shown standing smiling incorporating wearing an wearing harnehandle fedora with smile yellow sunglasses yellow sunglasses and people and street wth foreground with people people among people cars multiple billboards cityscape and street billboards asphalt street cute screenshot street”

What is unusual about this image?

It is unusual to see a corgi riding a bicycle, especially in a busy city like New York...

More examples

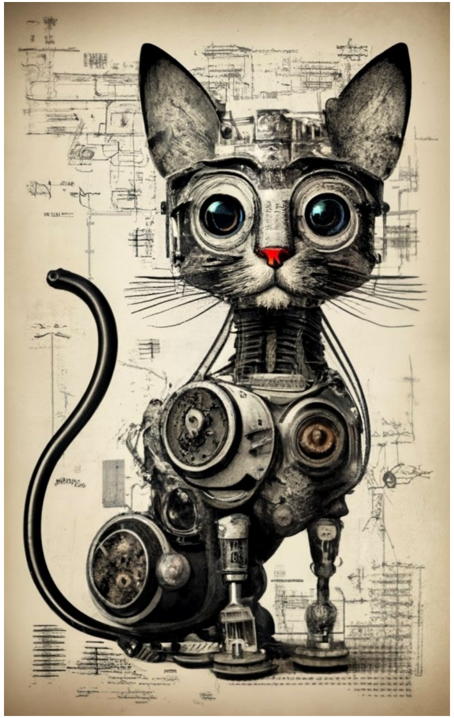
Photos



a colcord vindic analogue closeup of yellow bus mirror through a mirror mirror mirror depicting a yellow vehicle it on black bicycle foreground behind a black pol handle a boards behind blur towards a white boards boards above background silhousky sky white siding and poles pole besides wooden pole lb orange green green handle reflection content reflection black black round mirror rotterdam wisconsin congestion vehicle mirror mirror vintage hdr reflection glass mirror

More examples

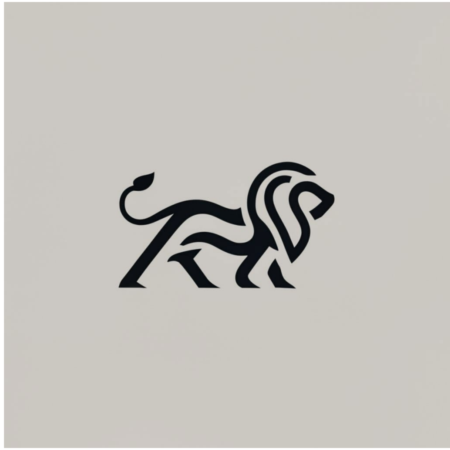
Creative Arts



an hdr painting robotic cat aus blackandwhite steampunk cyborg cat with eye with crook standing an on black leash robot steampunk steampunk os steampunk punk robot resemb exhibiting called an filly cat shown standing neck showcasing relating an derel steampunk mechanism robot with optom eye optomeyes optomelenses silver monochrome brown monochrome consist overlooking with derel mechanism overlooking scratched cities faded codes on beige yellow background portrait eyebrow steampunk steampunk artwork

More examples

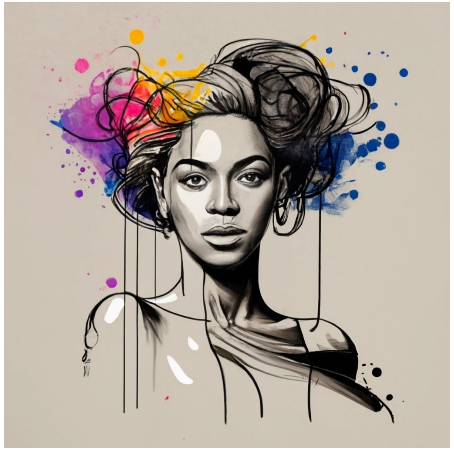
Abstract Symbols



a coleman vscocam appreciation an lions animal walking side winding winding lion on a beige pared it minimalist profile silhouette featuring featuring a black pol curled a silhouette atop lineup in a atmospheric silhouette silhouette on pale beige pared backgrounds grey background minimalist minimalist gladly featuring minimalist silhouette black and blk navy black white symbol modern i modernist minimalist white lions lions atx wsj fintech minimalist line symbol render ometric silhouette profile symbol

More examples

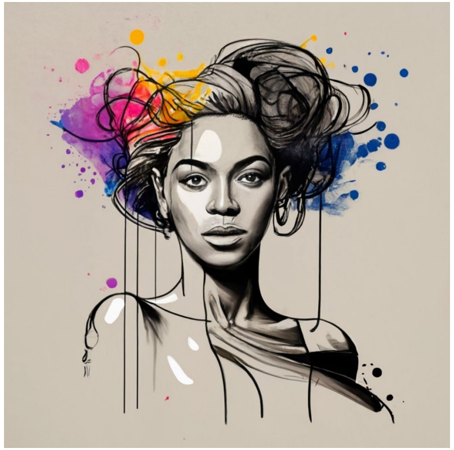
Factual Knowledge



a colalbugmbearts drawing beyonce beyonce portrait woman portrait with swept swept curls in a silver dress it with colourful watercolor there with a a colorful pollouda cloud above atop with a colorful watercolor graffiti atop beige beige background shadows grey background with spots wth with woman ear bold magenta purple black sleek off sleeveless black and white black and white yellow swoo curls rihanna jimi supermodel abstract inktober drawing illustration stration face eyebrow portrait

More examples

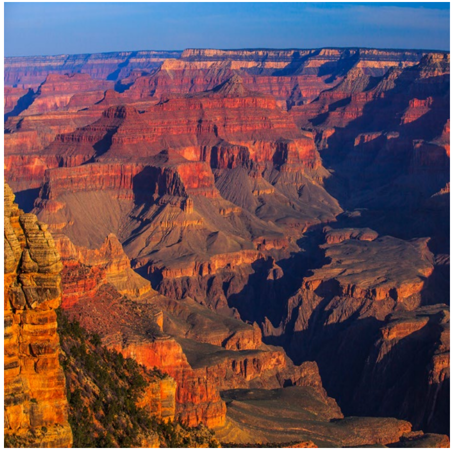
Factual Knowledge



a colalbugmbearts drawing beyonce beyonce portrait woman portrait with swept swept curls in a silver dress it with colourful watercolor there with a a colorful pollouda cloud above atop with a colorful watercolor graffiti atop beige beige background shadows grey background with spots wth with woman ear bold magenta purple black sleek off sleeveless black and white black and white yellow swoo curls rihanna jimi supermodel abstract inktober drawing illustration stration face eyebrow portrait

More examples

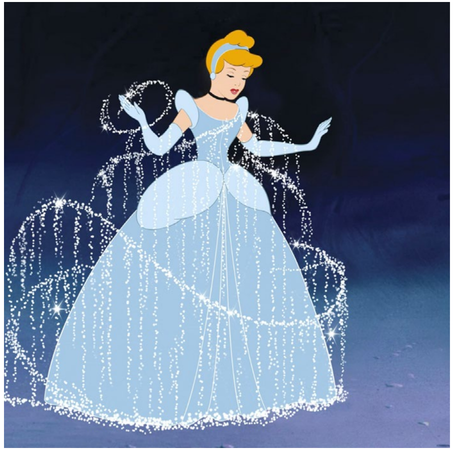
Factual Knowledge



a typical grand canyon many brown vast peaks mountains and canyon and overview surrounded an overview mostly overview canyon grand grand grand canyon canyon shown foreground note an overview canyon shown overview lineup whilst towards an morning morning dusk overview also morning morning blue sky and sky redness orange brown purple wth numerous towards vast cliffs numerous grand canyon grand cliffs mountains mostly blue background dusk overview grand grand canyon

More examples

Factual Knowledge



a illustration cartoon cinderella cinderella female fing examining offering sparkling webs wearing a pale dress it distributed sparkly strings overwhelmed distributed a a silver pol strings a string surrounding seen against a gray surface background right darkness purple background shadows snow ground darkness tree featuring wearing blue gloves pale pale blue white white dressed mostly silhouette yellow yellow hair ponytail disney rodgers imation animation cinderella gown illustration white face hime gown