Correct Captions are All you Need

The worth of metadata and correct caption

AngelBottomless, aria1th@Github
Contact : aria1th@naver.com
Yonsei University, department of AI

Index

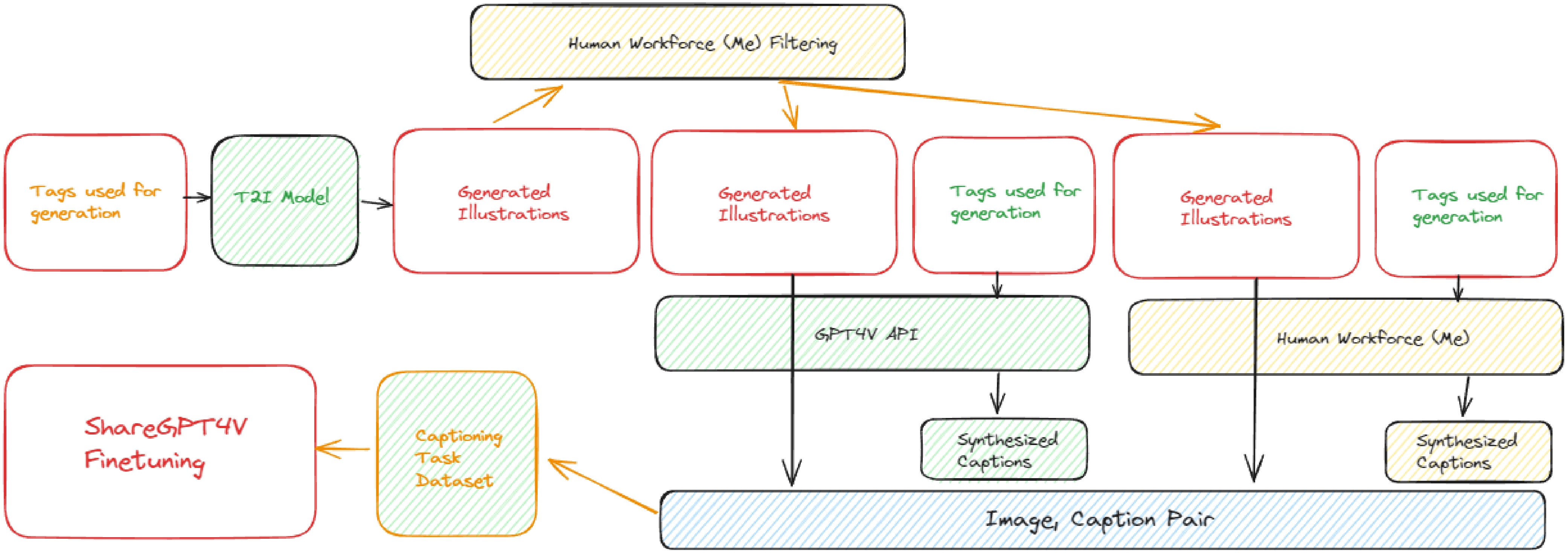
- Overview
- Theorical Background
  - Refined Dataset is all you need
  - Textbooks are all you need I, II & LIMA
  - DALLE-3, Principled Recaptioning : A Picture is Worth a Thousand Words
  - Densely Captioned Images
  - DinoV2, self supervised learning for image understanding
- Synthesizing Captions
  - Dataset selection
  - Model-wise results
  - Refining captions
- Future works

Current State of Image Generation

- SD
  - SD 1.x
    - Targeted 512x512 size image, finetuned without augmentation
    - Does not natively support various ratio / high resolution images (3-stage)
  - SD 2.1 (deprecated)
    - Consideration of various ratio
  - SD XL
    - Multi-stage decode-encode (Far better handling for various structures)
- Non-CLIP based (T5) models
  - Wuerstchen, Kandinsky, Pixart-alpha
- Others
  - DALLE-3, MJ…

Project Overview

- Purpose : Synthesize Illustration style caption



Correct Captions are All you Need

- Google, OpenAI has proven that **correct informative captions** are critical for better generative models.
- Microsoft has proven that **refined high quality dataset** is key for better models.

- SD XL has shown that **data augmentation** is necessary for various image generation.

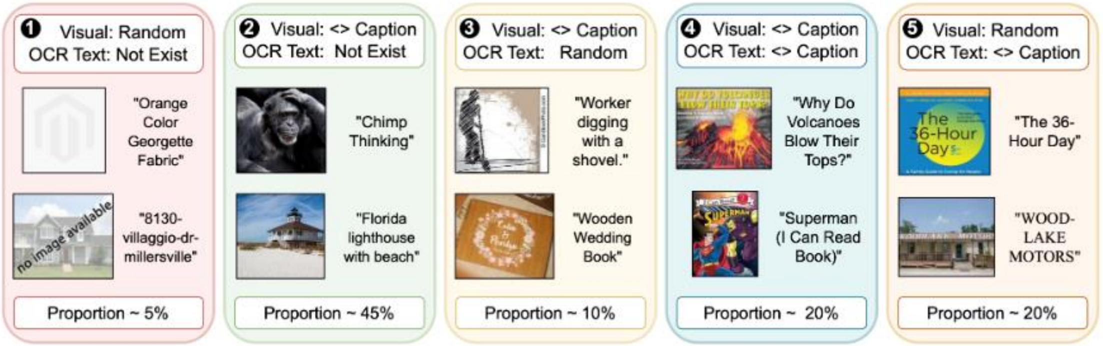
- We need better captions for better image/video generation.
- Here, we focus on finetuning on Illustration type dataset.

Why captioning is that important?

Refined Dataset is all you need

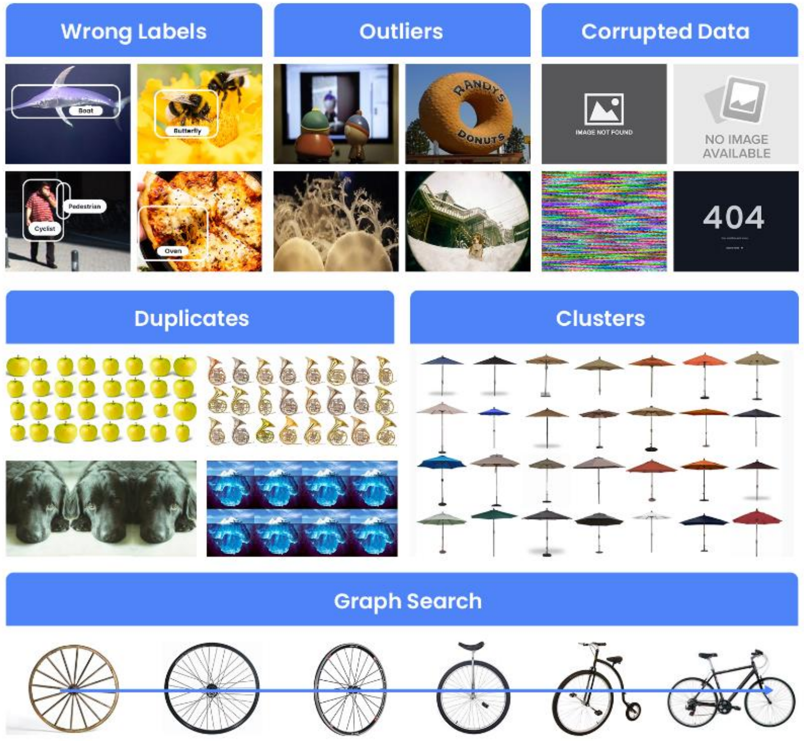
- **Noisy** nature of open dataset
  - Duplicates, Outliers, Mislabeling

A Look at the LAION Dataset

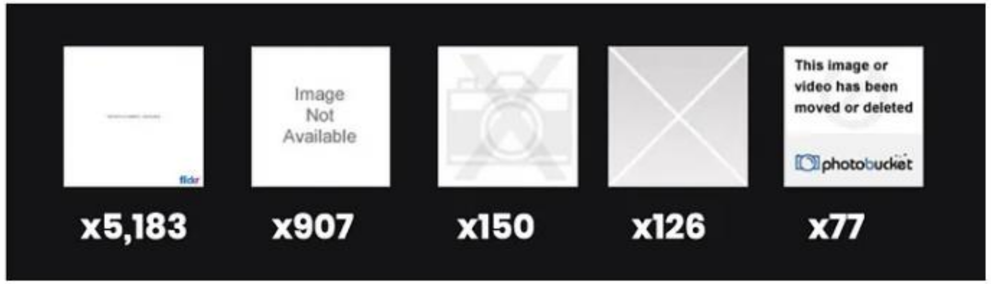


1. Our analysis shows an interesting observation: a large fraction of image-caption web datasets (such as LAION) have images that contain text inside them. Often, the text is the only feature correlated with the caption.
2. We aim to remove such images, as they promote model to learn optical character recognition rather than learning better visual features.

https://tmars-clip.github.io/



https://github.com/visual-layer/fastdup



Example classes of broken images that have valid labels. Many more such classes exist.

Textbooks are all you need, Microsoft

- Synthetic High-quality Dataset is **Really** important.

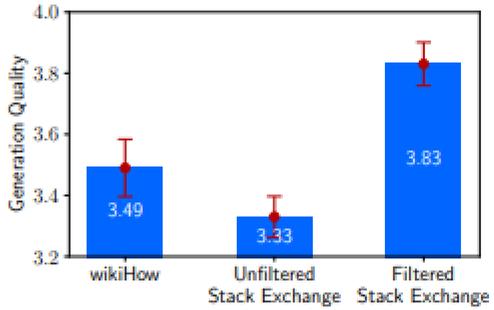


Figure 5: Performance of 7B models trained with 2,000 examples from different sources. **Filtered Stack Exchange** contains diverse prompts and high quality responses; **Unfiltered Stack Exchange** is diverse, but does not have any quality filters; **wikiHow** has high quality responses, but all of its prompts are "how to" questions.

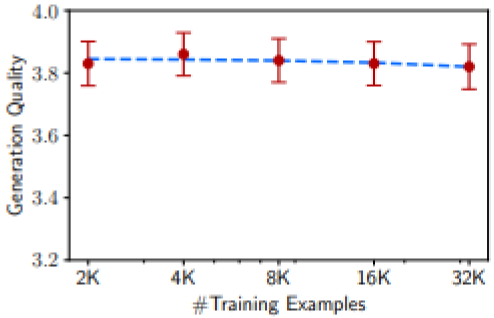
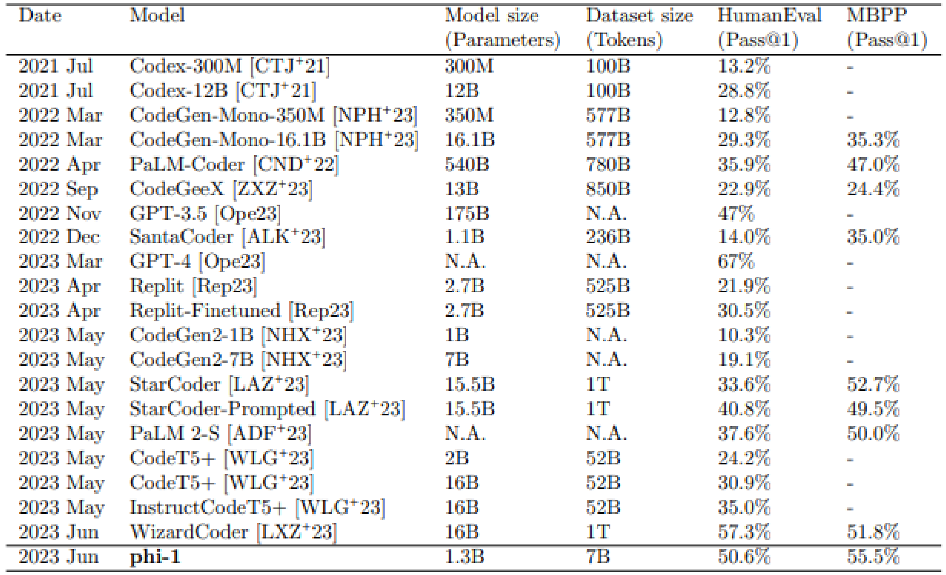


Figure 6: Performance of 7B models trained with exponentially increasing amounts of data, sampled from (quality-filtered) Stack Exchange. Despite an up to 16-fold increase in data size, performance as measured by ChatGPT plateaus.

LIMA, Figure 5. https://arxiv.org/pdf/2305.11206.pdf



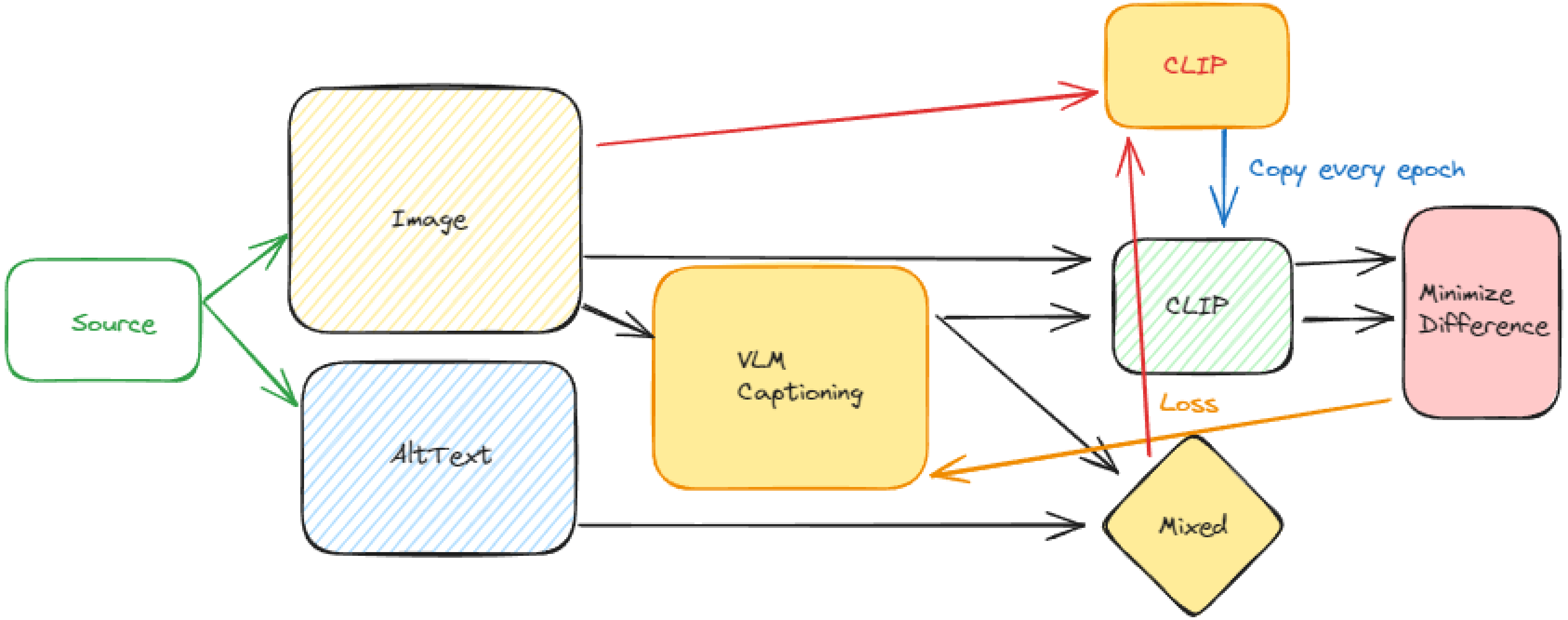
Date	Model	Model size (Parameters)	Dataset size (Tokens)	HumanEval (Pass@1)	MBPP (Pass@1)
2021 Jul	Codex-300M [CTJ$^+$21]	300M	100B	13.2%	-
2021 Jul	Codex-12B [CTJ$^+$21]	12B	100B	28.8%	-
2022 Mar	CodeGen-Mono-350M [NPH$^+$23]	350M	577B	12.8%	-
2022 Mar	CodeGen-Mono-16.1B [NPH$^+$23]	16.1B	577B	29.3%	35.3%
2022 Apr	PaLM-Coder [CND$^+$22]	540B	780B	35.9%	47.0%
2022 Sep	CodeGeeX [ZXZ$^+$23]	13B	850B	22.9%	24.4%
2022 Nov	GPT-3.5 [Ope23]	175B	N.A.	47%	-
2022 Dec	SantaCoder [ALK$^+$23]	1.1B	236B	14.0%	35.0%
2023 Mar	GPT-4 [Ope23]	N.A.	N.A.	67%	-
2023 Apr	Replit [Rep23]	2.7B	525B	21.9%	-
2023 Apr	Replit-Finetuned [Rep23]	2.7B	525B	30.5%	-
2023 May	CodeGen2-1B [NHX$^+$23]	1B	N.A.	10.3%	-
2023 May	CodeGen2-7B [NHX$^+$23]	7B	N.A.	19.1%	-
2023 May	StarCoder [LAZ$^+$23]	15.5B	1T	33.6%	52.7%
2023 May	StarCoder-Prompted [LAZ$^+$23]	15.5B	1T	40.8%	49.5%
2023 May	PaLM 2-S [ADF$^+$23]	N.A.	N.A.	37.6%	50.0%
2023 May	CodeT5+ [WLG$^+$23]	2B	52B	24.2%	-
2023 May	CodeT5+ [WLG$^+$23]	16B	52B	30.9%	-
2023 May	InstructCodeT5+ [WLG$^+$23]	16B	52B	35.0%	-
2023 Jun	WizardCoder [LXZ$^+$23]	16B	1T	57.3%	51.8%
2023 Jun	**phi-1**	1.3B	7B	50.6%	55.5%

Table 1: We use self-reported scores whenever available. Despite being trained at vastly smaller scale, **phi-1** outperforms competing models on HumanEval and MBPP, except for GPT-4 (also WizardCoder obtains better HumanEval but worse MBPP).

Textbooks are all you need, Table 1. https://arxiv.org/abs/2306.11644

DALLE-3 - Not Directly About Image Generation

- “Improving Image Generation with Better Captions”
  - https://cdn.openai.com/papers/dall-e-3.pdf
- **Minimizing** the gap of LLM and output and CLIP output*



Key idea of Re-Captioning in DALLE-3

Principled Recaptioning - Google

- Short Alignment (300 step) of PaLI for generating captions
- **Small(????)** 10M Subset recaptioned for SD 1.4
- Observed semantic improvements with **long synthesized captions**

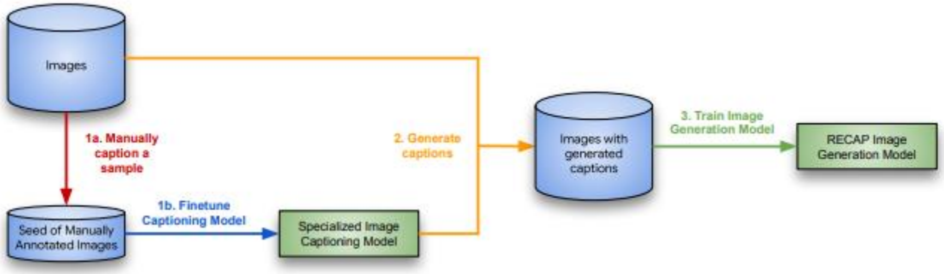

Images
1a. Manually caption a sample
Seed of Manually Annotated Images
1b. Finetune Captioning Model
Specialized Image Captioning Model
2. Generate captions
Images with generated captions
3. Train Image Generation Model
RECAP Image Generation Model

Figure 2. Schematic diagram of our method RECAP. In steps (1a) and (1b) we fine-tune an image-to-text captioning model on a small set of detailed human captions. In step (2) we use this fine-tuned model to recaption the images in the training dataset of a text-to-image model, and with this dataset, in step (3) we train an image generation model with the recaptioned dataset.

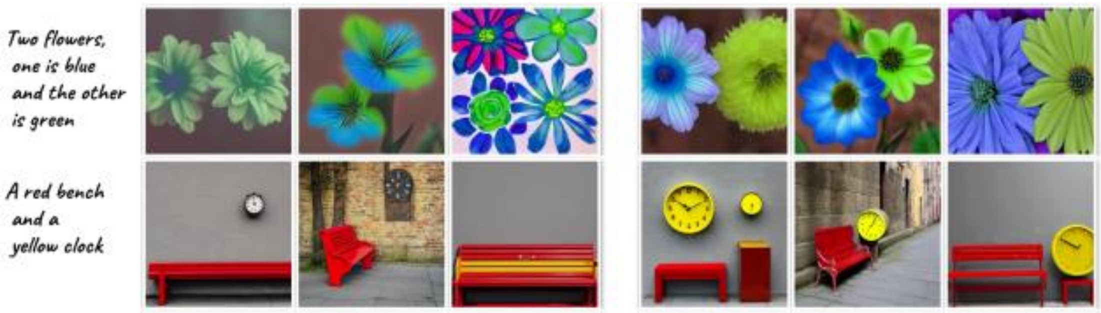
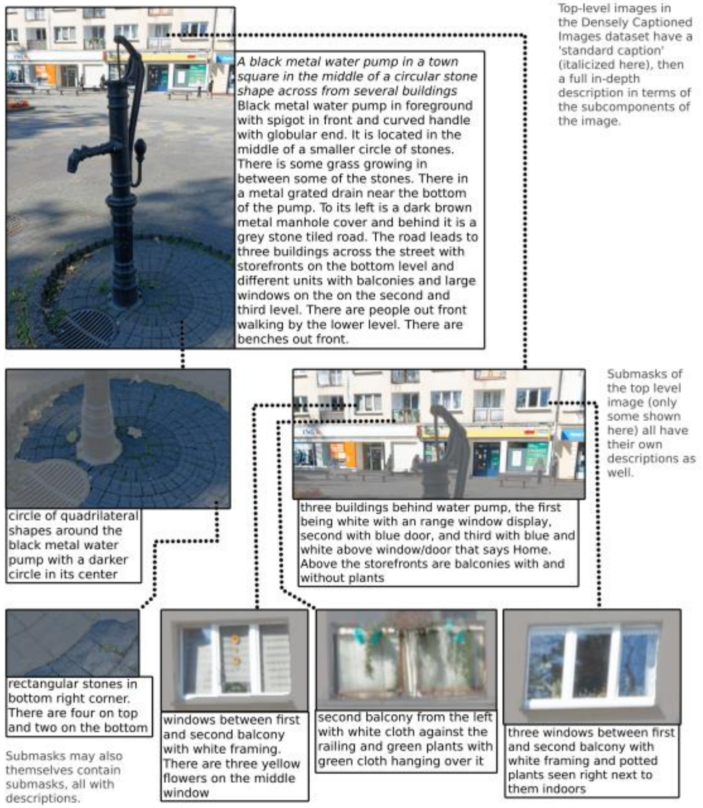

Two flowers, one is blue and the other is green
A red bench and a yellow clock

Fig.1 and Fig.2
https://arxiv.org/pdf/2310.16656.pdf

Densely Captioned Images, Meta

- **Hierarchical** Captioning
  - The one CLIP has failed before
- Strict Quality Check
  - Acknowledgement of cleaned dataset importance



A black metal water pump in a town square in the middle of a circular stone shape across from several buildings Black metal water pump in foreground with spigot in front and curved handle with globular end. It is located in the middle of a smaller circle of stones. There is some grass growing in between some of the stones. There in a metal grated drain near the bottom of the pump. To its left is a dark brown metal manhole cover and behind it is a grey stone tiled road. The road leads to three buildings across the street with storefronts on the bottom level and different units with balconies and large windows on the on the second and third level. There are people out front walking by the lower level. There are benches out front.

Top-level images in the Densely Captioned Images dataset have a 'standard caption' (italicized here), then a full in-depth description in terms of the subcomponents of the image.

circle of quadrilateral shapes around the black metal water pump with a darker circle in its center

three buildings behind water pump, the first being white with an range window display, second with blue door, and third with blue and white above window/door that says Home. Above the storefronts are balconies with and without plants

Submasks of the top level image (only some shown here) all have their own descriptions as well.

rectangular stones in bottom right corner. There are four on top and two on the bottom

windows between first and second balcony with white framing. There are three yellow flowers on the middle window

second balcony from the left with white cloth against the railing and green plants with green cloth hanging over it

three windows between first and second balcony with white framing and potted plants seen right next to them indoors

Submasks may also themselves contain submasks, all with descriptions.

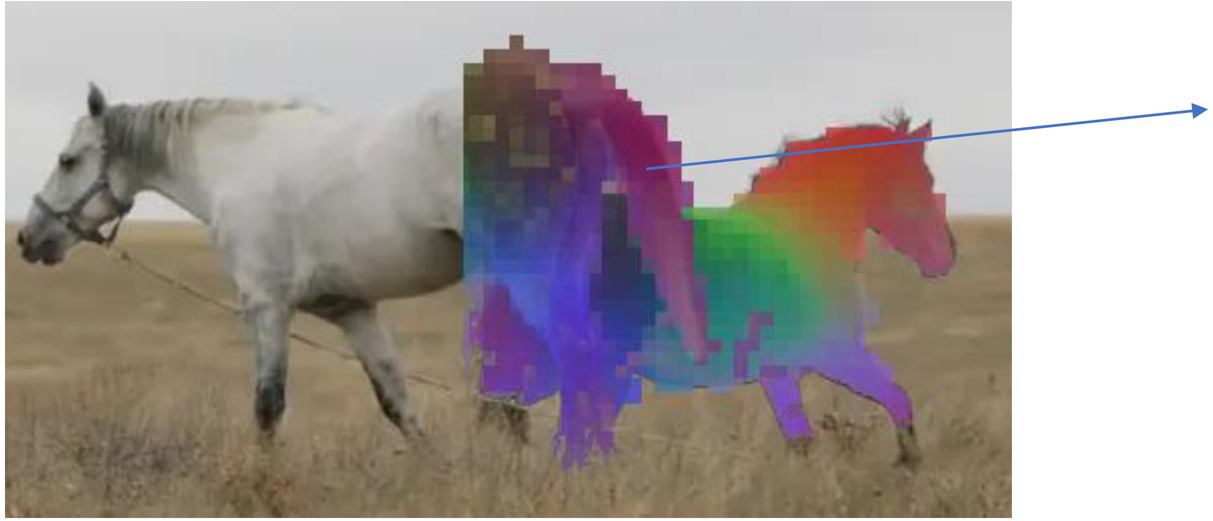
Figure 1 One example from the Densely Captioned Images dataset. Only part of the submask hierarchy is shown.

https://github.com/facebookresearch/DCI

Extra. DinoV2 – Captions are Flawed

- Dinov2 abandoned captions, due to lacking dense information.
- Instead, it used Self-Supervised-Learning with patches.

- Q. Can we use it again for captioning?



“Horse tail”?

Why Illustration?

- General dataset
  - Photo / logo based dataset(LAION, COCO etc) had large rework
  - *Especially aesthetic photo (DCI, 2023)
- Illustrations, requires dense captions too
- VLM-based recaptioning requires:
  - Good alttext
  - Good VLM
- VLM is commonly supported with CLIP Retrieval
  - CLIP does **not** do well with Illustrations

ONKOLOJİDE
YARINLARIN FAR
ASTRAZE

Astra Zeneca Oncology launch design & CGI

Olivier olive tree

→ Synthesize caption to densely annotate illustrations

https://rom1504.github.io/clip-retrieval/?back=https%3A%2F%2Fknn.laion.ai&index=laion5B-H-14&useMclip=false

Synthesizing captions

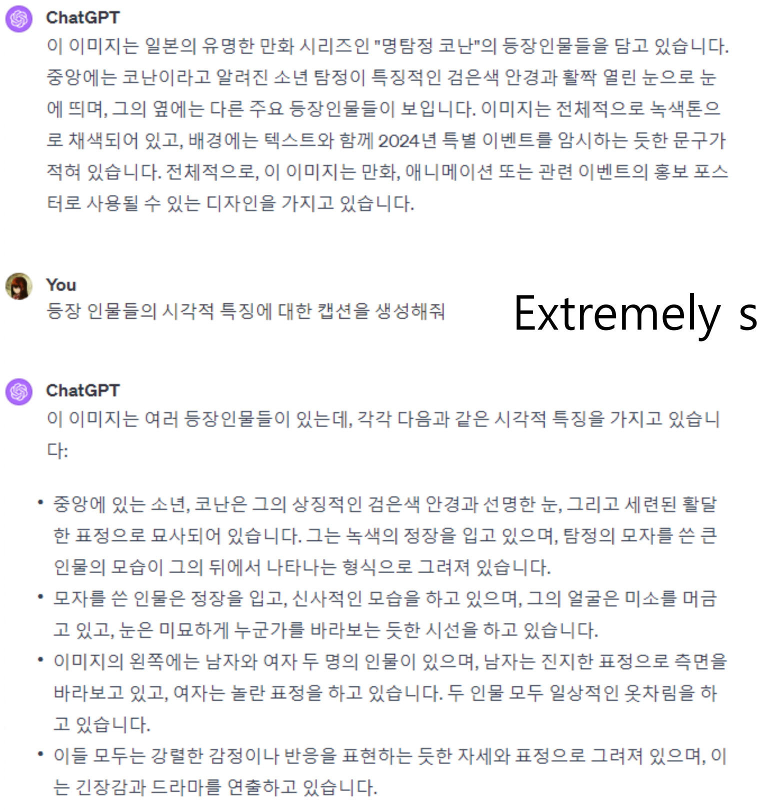
Tag-Retrieval / Normal captioning

Current State of Captioning

- GPT-4V
  - ShareGPT4V, CapsFusion Dataset, State of the art
- LLaVa / VLMs (ShareGPT4V, Fuyu)
  - Heavy, reasonable performance, slow
- BLIP Captioning
  - Fast, low performance, light
- Vocabulary-based-Captioning
  - Fast, **limited vocabulary**, thresholding problem
- No Captions (source)
  - Known as the worst way to use the dataset but no additional cost

Usability

- GPT-4V
    - Definitely
- ShareGPT4V
    - Photography – Very well
    - Logo – Not likely
    - Illustrations – Misses all other stuff
- BLIP Captioning
    - Unusable
- Vocabulary-based-Captioning
    - Usable, too many false positives
- Gemini Pro Vision
    - 1+shot - **Usable**



ChatGPT

이 이미지는 일본의 유명한 만화 시리즈인 "명탐정 코난"의 등장인물들을 담고 있습니다. 중앙에는 코난이라고 알려진 소년 탐정이 특징적인 검은색 안경과 활짝 열린 눈으로 눈에 띄며, 그의 옆에는 다른 주요 등장인물들이 보입니다. 이미지는 전체적으로 녹색톤으로 채색되어 있고, 배경에는 텍스트와 함께 2024년 특별 이벤트를 암시하는 듯한 문구가 적혀 있습니다. 전체적으로, 이 이미지는 만화, 애니메이션 또는 관련 이벤트의 홍보 포스터로 사용될 수 있는 디자인을 가지고 있습니다.

You

등장 인물들의 시각적 특징에 대한 캡션을 생성해줘

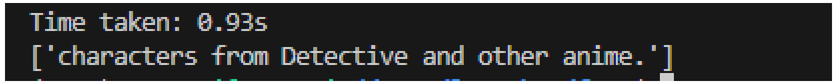
ChatGPT

이 이미지는 여러 등장인물들이 있는데, 각각 다음과 같은 시각적 특징을 가지고 있습니다:

- 중앙에 있는 소년, 코난은 그의 상징적인 검은색 안경과 선명한 눈, 그리고 세련된 활달한 표정으로 묘사되어 있습니다. 그는 녹색의 정장을 입고 있으며, 탐정의 모자를 쓴 큰 인물의 모습이 그의 뒤에서 나타나는 형식으로 그려져 있습니다.
- 모자를 쓴 인물은 정장을 입고, 신사적인 모습을 하고 있으며, 그의 얼굴은 미소를 머금고 있고, 눈은 미묘하게 누군가를 바라보는 듯한 시선을 하고 있습니다.
- 이미지의 왼쪽에는 남자와 여자 두 명의 인물이 있으며, 남자는 진지한 표정으로 측면을 바라보고 있고, 여자는 놀란 표정을 하고 있습니다. 두 인물 모두 일상적인 옷차림을 하고 있습니다.
- 이들 모두는 강렬한 감정이나 반응을 표현하는 듯한 자세와 표정으로 그려져 있으며, 이는 긴장감과 드라마를 연출하고 있습니다.

Extremely smart

Fuyu tried too
At least it tried



```
Time taken: 0.93s
['characters from Detective and other anime.']
```

Let's not talk about BLIP

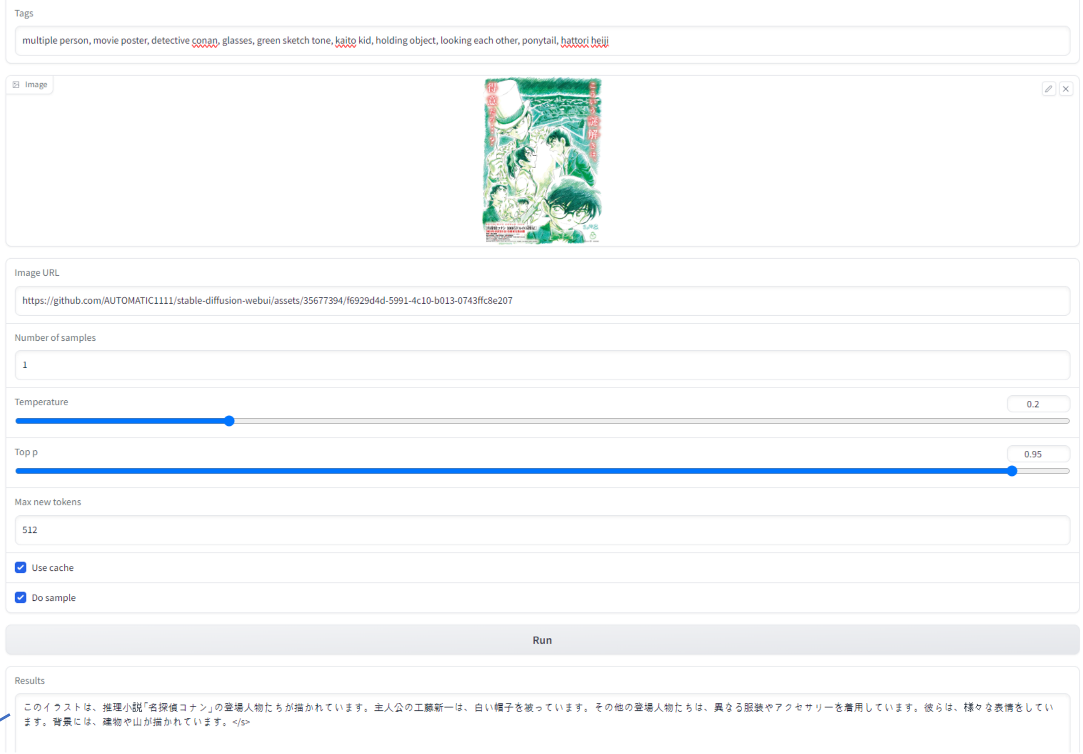
Sources for Datasets

- Danbooru
  - Tagged large database for **illustration-types**
  - Fair-use available, opt-out*, ~7M
- Unsplash
  - Tagged large **photo style** database
  - Fair-use available, 4.8M+
- AIBooru
  - “Generated image” database
  - Illlustrations (98.4%), Photorealistic images (1.5%)
  - Tagged with metadata and user tags
  - 5.3k images(23-12-17)

Synthesizing captions for **Illustration Types**

- LAION, Flickr30k, Unsplash25k dataset are biased for **real images**
- ShareGPT4V, BLIP has low performance on **illustrations**
  - Mainly due to dataset bias
- GPT4V is expensive and not local
  - 7$ / 1k images (minimal)

→ Finetune ShareGPT4V for Illustrations



Tags: multiple person, movie poster, detective conan, glasses, green sketch tone, kaito kid, holding object, looking each other, ponytail, hattori heiji

Image URL: https://github.com/AUTOMATIC1111/stable-diffusion-webui/assets/35677394/f6929d4d-5991-4c10-b013-0743ffc8e207

Number of samples: 1

Temperature: 0.2

Top p: 0.95

Max new tokens: 512

☑ Use cache

☑ Do sample

Results:

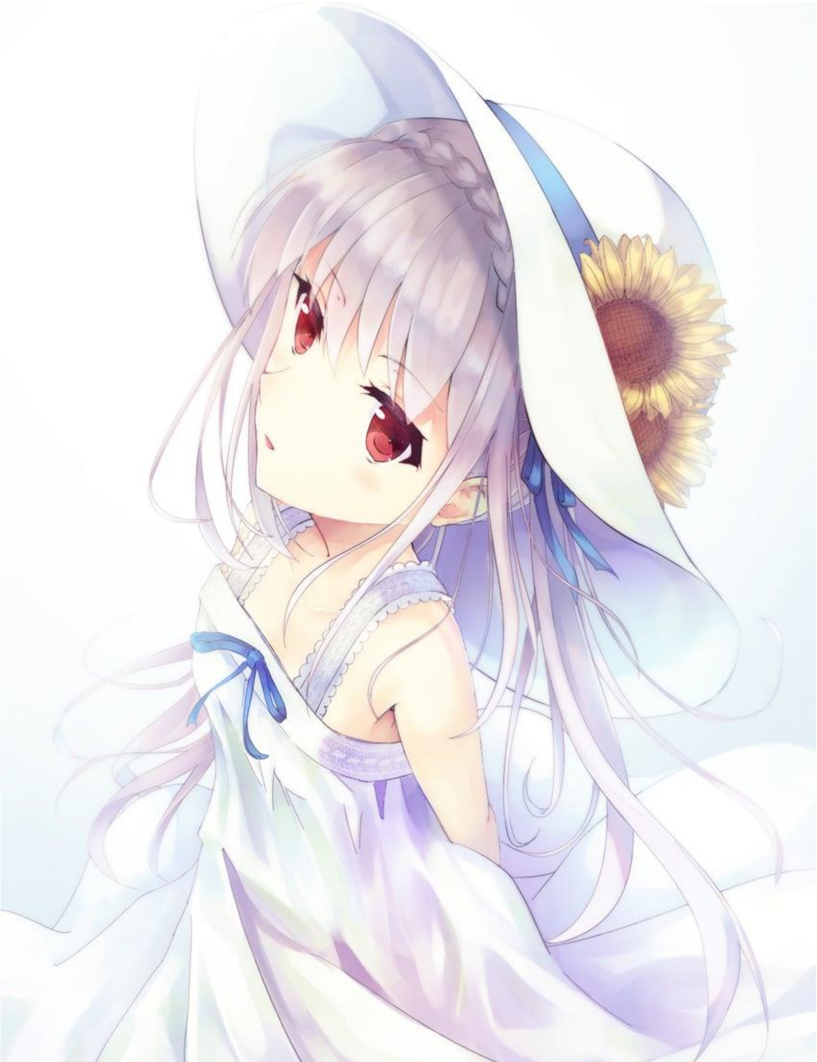
このイラストは、推理小説「名探偵コナン」の登場人物たちが描かれています。主人公の工藤新一は、白い帽子を被っています。その他の登場人物たちは、異なる服装やアクセサリーを着用しています。彼らは、様々な表情をしています。背景には、建物や山が描かれています。</s>

Breaking ShareGPT4V with “detective conan’ movie illustration

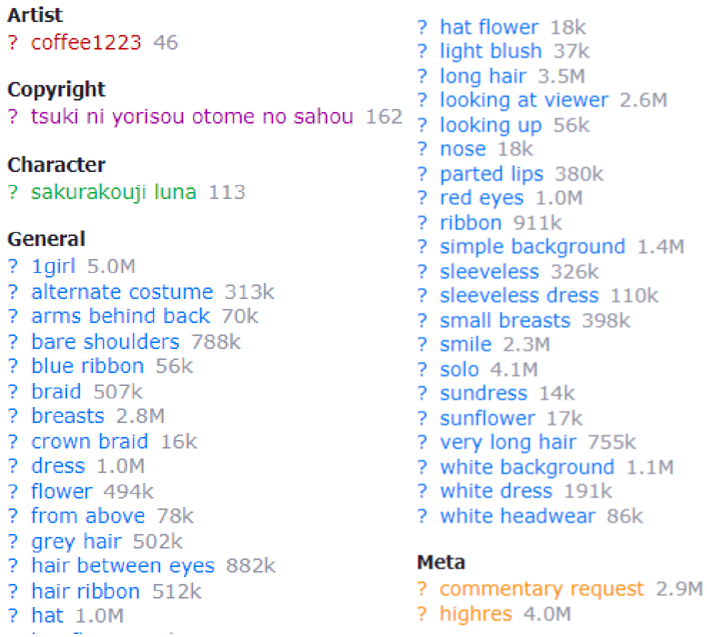
Speaking Japanese because poster seems to be Japanese
“Kudo Shinichi is wearing white hat (WRONG)
Others are wearing some accessories (WHAT)
There are buildings or mountain in background (NO)

Performance test examples

ShareGPT4V



- "The image portrays a young girl with long, flowing **blonde hair**. Her eyes, a striking shade of red, stand out against her fair complexion. She is dressed in a white dress adorned with blue ribbons, adding a touch of elegance to her appearance. On her head, she wears a white hat, which is embellished with a **single** yellow sunflower, adding a pop of color to the otherwise monochromatic ensemble. The background is a light blue gradient, providing a serene and calm atmosphere that complements the girl's peaceful demeanor. The image does not contain any discernible text or additional objects. The girl is the central figure, **and her position relative to the background suggests she is standing in front of it.** The image does not provide any information about the girl's actions, as she appears to be in a state of stillness. The image is a digital illustration, and all descriptions are based on the visible content of the image."



Artist
? coffee1223 46

Copyright
? tsuki ni yorisou otome no sahou 162

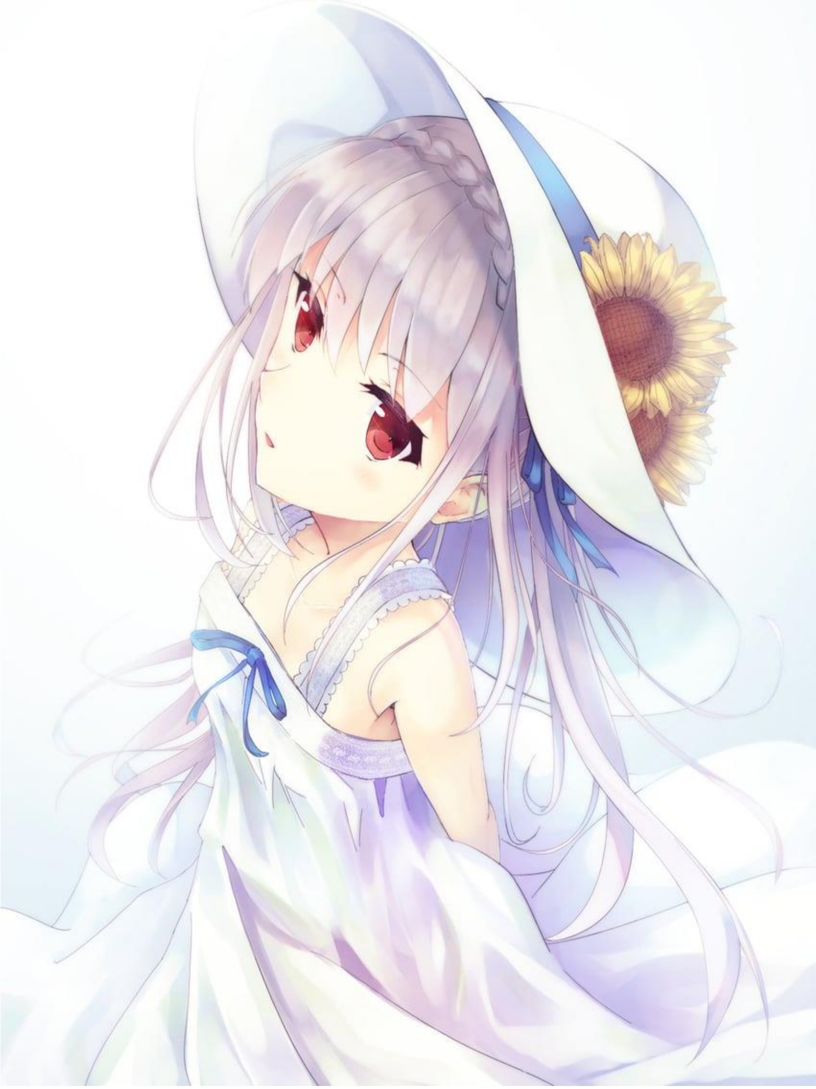
Character
? sakurakouji luna 113

General
? 1girl 5.0M
? alternate costume 313k
? arms behind back 70k
? bare shoulders 788k
? blue ribbon 56k
? braid 507k
? breasts 2.8M
? crown braid 16k
? dress 1.0M
? flower 494k
? from above 78k
? grey hair 502k
? hair between eyes 882k
? hair ribbon 512k
? hat 1.0M
? hat flower 18k
? light blush 37k
? long hair 3.5M
? looking at viewer 2.6M
? looking up 56k
? nose 18k
? parted lips 380k
? red eyes 1.0M
? ribbon 911k
? simple background 1.4M
? sleeveless 326k
? sleeveless dress 110k
? small breasts 398k
? smile 2.3M
? solo 4.1M
? sundress 14k
? sunflower 17k
? very long hair 755k
? white background 1.1M
? white dress 191k
? white headwear 86k

Meta
? commentary request 2.9M
? highres 4.0M

https://cdn.donmai.us/sample/45/37/__sakurakouji_luna_tsuki_ni_yorisou_otome_no_sahou_drawn_by_coffee1223__sample-4537da0289a2adaa72fd135c471bd011.jpg

GPT-4V

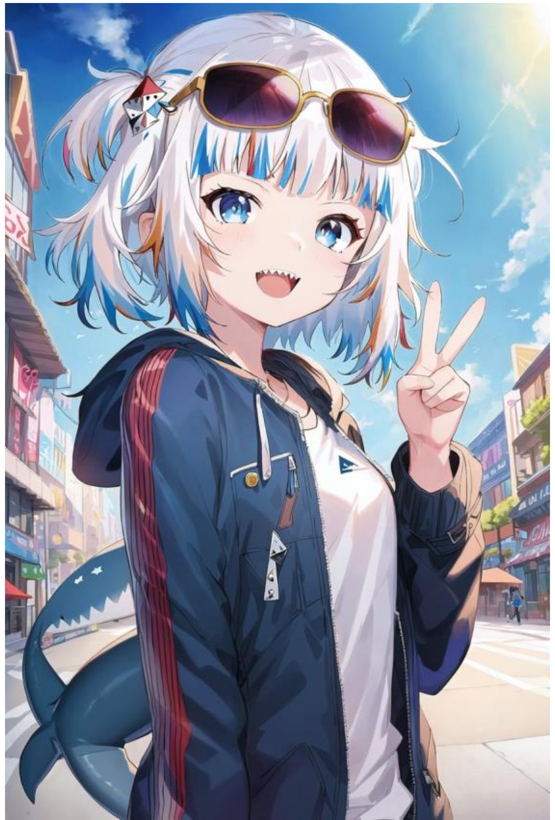


- "This illustration depicts a **single female character**, Sakurakouji Luna from the copyright series 'Tsuki ni Yorisou Otome no Sahou'. The artist is identified as Coffee1223. The character is portrayed in an alternate costume, with her arms positioned behind her back, and is wearing a white, sleeveless dress. She has bare shoulders and small breasts are implied. Her attire includes a blue ribbon, and she sports a crown braid with very long, grey hair that includes a braid. There is a flower – specifically a sunflower – attached to her white headwear, which resembles a hat. The character has red eyes, light blush on her cheeks, and her hair falls between her eyes. She is looking at the viewer with a subtle smile and parted lips. The background of the image is simple and white, emphasizing the character. The dress appears to be a sundress, suitable for warm weather, and the character's overall expression and pose convey a gentle and approachable demeanor."

Obviously a several generation gap is there.

Gemini pro Vision

- It failed severely with 0-Shot.
- 1-Shot Result:

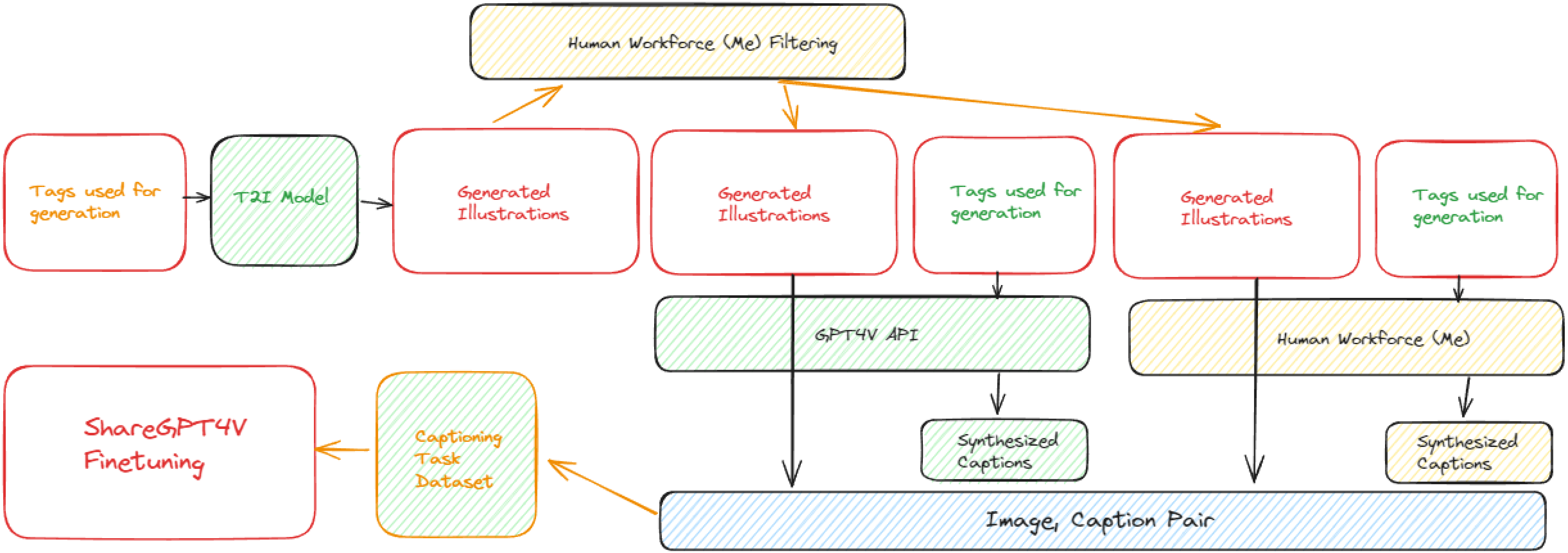


copyright: hololive hololive_english
character: gawr_gura
general tags: 1girl architecture bangs blue_eyes blue_sky blush building city cityscape cloud cloudy_sky day eyewear_on_head hair_ornament hood hoodie looking_at_viewer multicolored_hair open_mouth outdoors road sharp_teeth short_hair sidelocks sky smile solo standing streaked_hair street sun sunglasses tail teeth two_side_up v virtual_youtuber

Gawr Gura, a virtual YouTuber from Hololive English, is depicted in this image. She is the only character present (1girl). Her appearance includes blue_eyes, multicolored_hair, and sharp_teeth. She is wearing a hoodie and sunglasses, with her hair styled with sidelocks and a hair_ornament. The background shows a city street with buildings, and the sky is cloudy_sky with some sun peeking through. Gura is standing in the street, facing the viewer with a smile and making a peace sign.

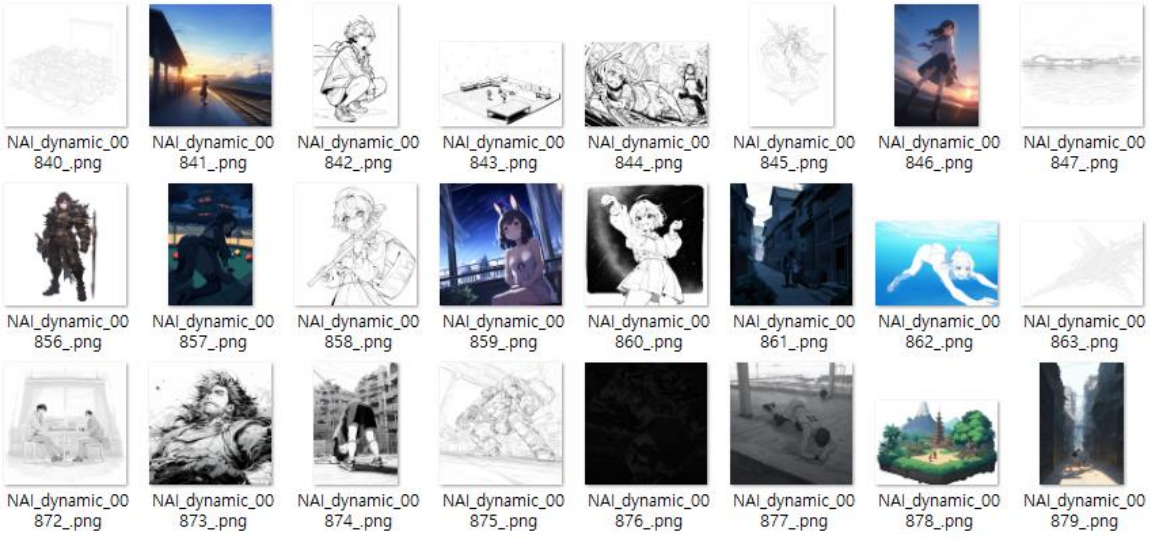
Some missing tags, but reasonable redundancy removal?

Pipeline Overview

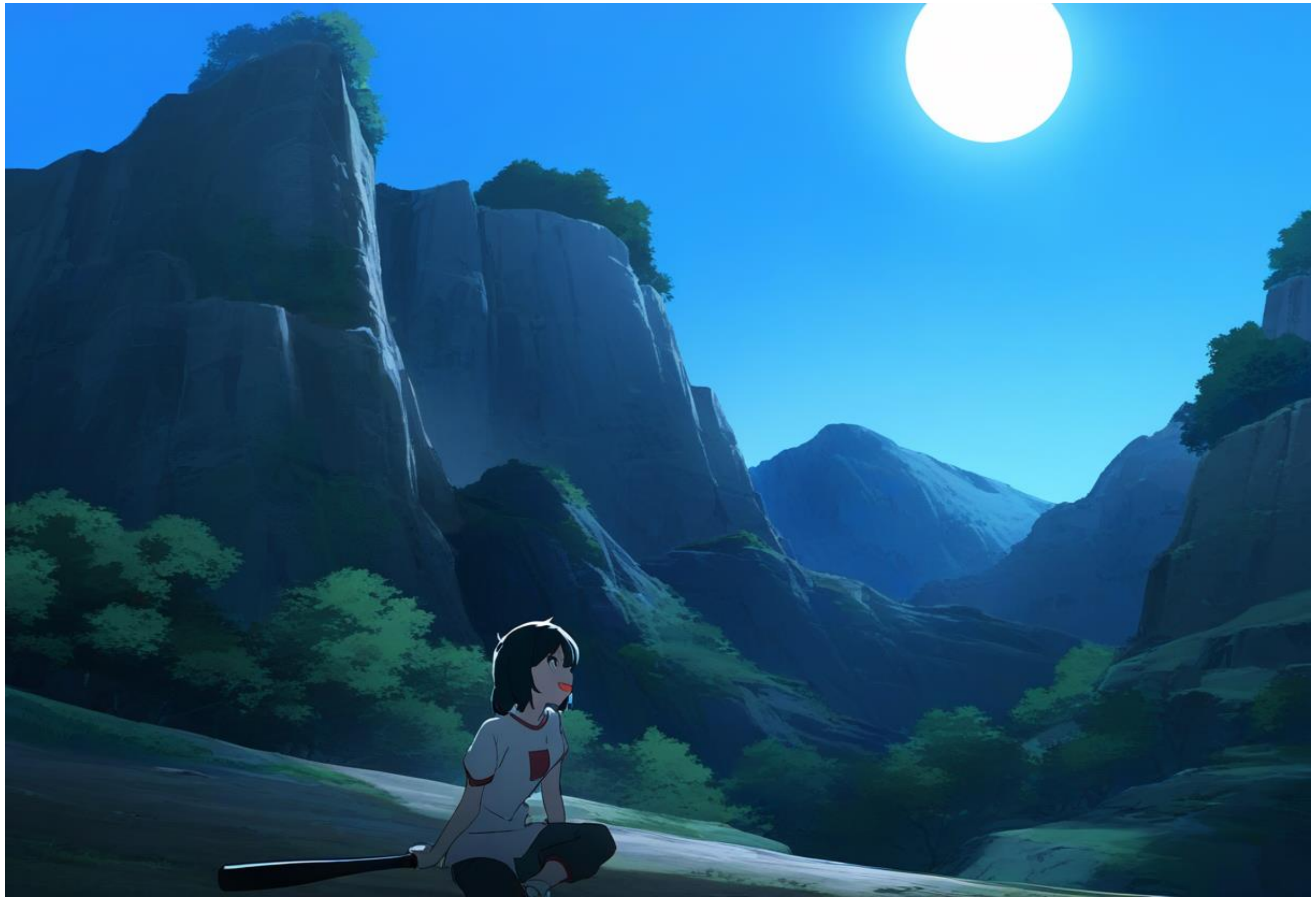


Sanity check - dataset

- Generated 1000 Text-image pair from Novel.AI
  - Filtered 230 pair, captioned without tag by GPT4V API
  - 20 Manually captioned images
- Chosen 200 text-image pair from aibooru dataset
  - Filtered 100 pair, captioned with tag by GPT4V API
  - 20 Manually captioned images

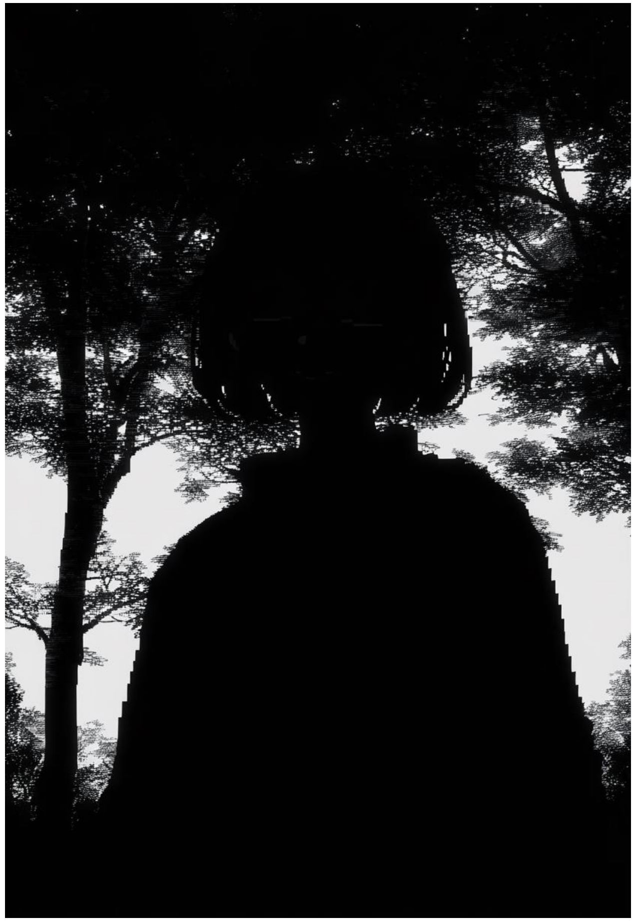


Examples of captioning (No tags)



The illustration is **far view** of nature and 1girl.
A girl with black hair, black eyes, short hair, is sitting on the rocky ground. She is wearing a white shirts with red collar. There is square like symbol drawn in white shirt. She is wearing a black jeans, and looking up the sky. She is holding a baseball bat on the ground.
Rocky mountains with few trees are located at background. Blue sky with shiny moon-like object is on the sky.

Failure Cases (No tags)



GPT-4V Response : NONE(Empty)

The illustration is drawn with extremely low exposure to light. There are Tree silhouette in background, with minimal light.
The short-hair girl is seemed to be located at the center of the illustration. It has rough boundary, which is possibly result of pixel art.
The image is composed with greyscale colors, it is upper body shot.

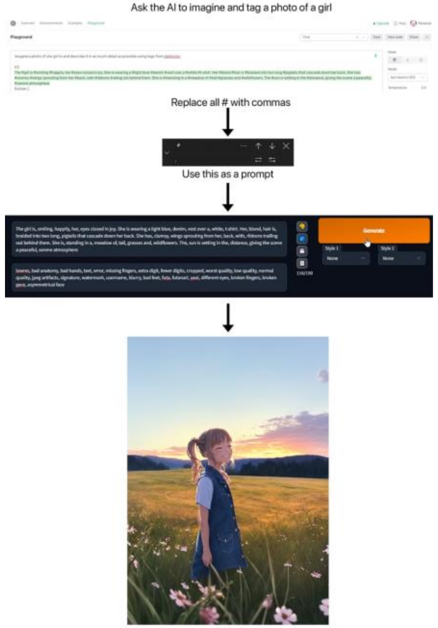
AIBooru Dataset

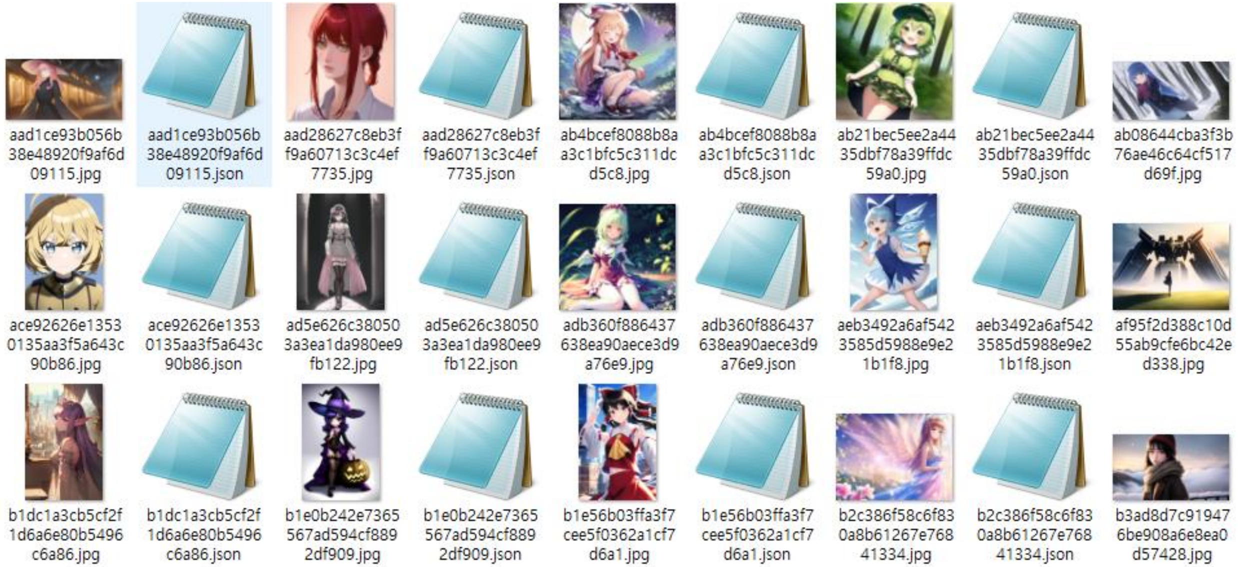
- Consistent Tagging and Captions
- Crowdsourced* refined dataset with quality check

- Does not mean **caption** is directly corresponding to Image
- But user-tags do correspond, can fix biases

Dataset Bias

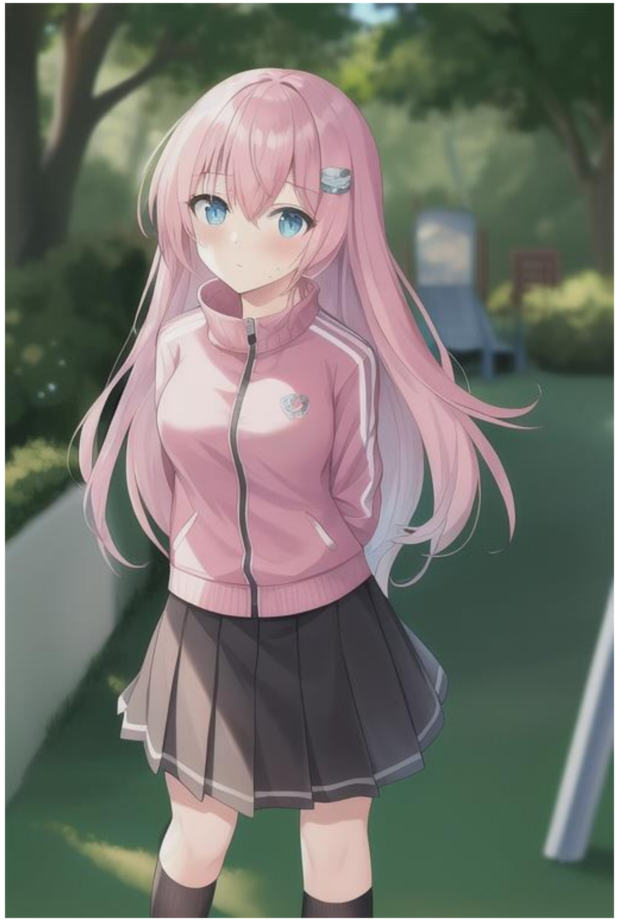
- 99% Portrait-style
- 90% female
- 2% multiple person
- 5% greyscale


Ask the AI to imagine and tag a photo of a girl
Replace all # with commas
Use this as a prompt


aad1ce93b056b38e48920f9af6d09115.jpg
aad1ce93b056b38e48920f9af6d09115.json
aad28627c8eb3ff9a60713c3c4ef7735.jpg
aad28627c8eb3ff9a60713c3c4ef7735.json
ab4bcef8088b8aa3c1bfc5c311dcd5c8.jpg
ab4bcef8088b8aa3c1bfc5c311dcd5c8.json
ab21bec5ee2a4435dbf78a39ffdc59a0.jpg
ab21bec5ee2a4435dbf78a39ffdc59a0.json
ab08644cba3f3b76ae46c64cf517d69f.jpg
ace92626e13530135aa3f5a643c90b86.jpg
ace92626e13530135aa3f5a643c90b86.json
ad5e626c380503a3ea1da980ee9fb122.jpg
ad5e626c380503a3ea1da980ee9fb122.json
adb360f886437638ea90aece3d9a76e9.jpg
adb360f886437638ea90aece3d9a76e9.json
aeb3492a6af5423585d5988e9e21b1f8.jpg
aeb3492a6af5423585d5988e9e21b1f8.json
af95f2d388c10d55ab9cfe6bc42ed338.jpg
b1dc1a3cb5cf2f1d6a6e80b5496c6a86.jpg
b1dc1a3cb5cf2f1d6a6e80b5496c6a86.json
b1e0b242e7365567ad594cf8892df909.jpg
b1e0b242e7365567ad594cf8892df909.json
b1e56b03ffa3f7cee5f0362a1cf7d6a1.jpg
b1e56b03ffa3f7cee5f0362a1cf7d6a1.json
b2c386f58c6f830a8b61267e76841334.jpg
b2c386f58c6f830a8b61267e76841334.json
b3ad8d7c919476be908a6e8ea0d57428.jpg

And still, data poisoning*

Example of captioning



copyright: bocchi_the_rock!
character: gotou_hitori
general tags: 1girl arms_behind_back bangs black_skirt black_socks blue_eyes blue_sky blurry blurry_background breasts closed_mouth cloud day dot_mouth foot_out_of_frame grass hair_between_eyes hair_ornament jacket kneehighs long_hair long_sleeves looking_at_viewer outdoors pink_hair pink_jacket pleated_skirt shiny shiny_hair sidelocks skirt sky socks solo standing track_jacket tree tree_shade

"The character Gotou Hitori from 'Bocchi the Rock!' is depicted as a 1girl, standing outdoors with a blue_sky and tree_shade in the blurry_background, which also includes subtle elements like cloud, grass, and a tree. She is wearing a pink_jacket, which is a track_jacket, paired with a black_skirt, in this case a pleated_skirt, and has her long_sleeves visible. Her long_hair is pink and exhibits a shiny_hair effect and includes details such as bangs, hair_between_eyes, and sidelocks, with a hair_ornament adding to her appearance. She has blue_eyes, a closed_mouth with a trademark dot_mouth, and is looking_at_viewer. Other features include black_socks, specifically kneehighs, and her arms_behind_back stance, with a footnote that her foot_out_of_frame.

Additional tags : shiny, solo, jacket, skirt, socks

Captioning Interface

Current image

Current index

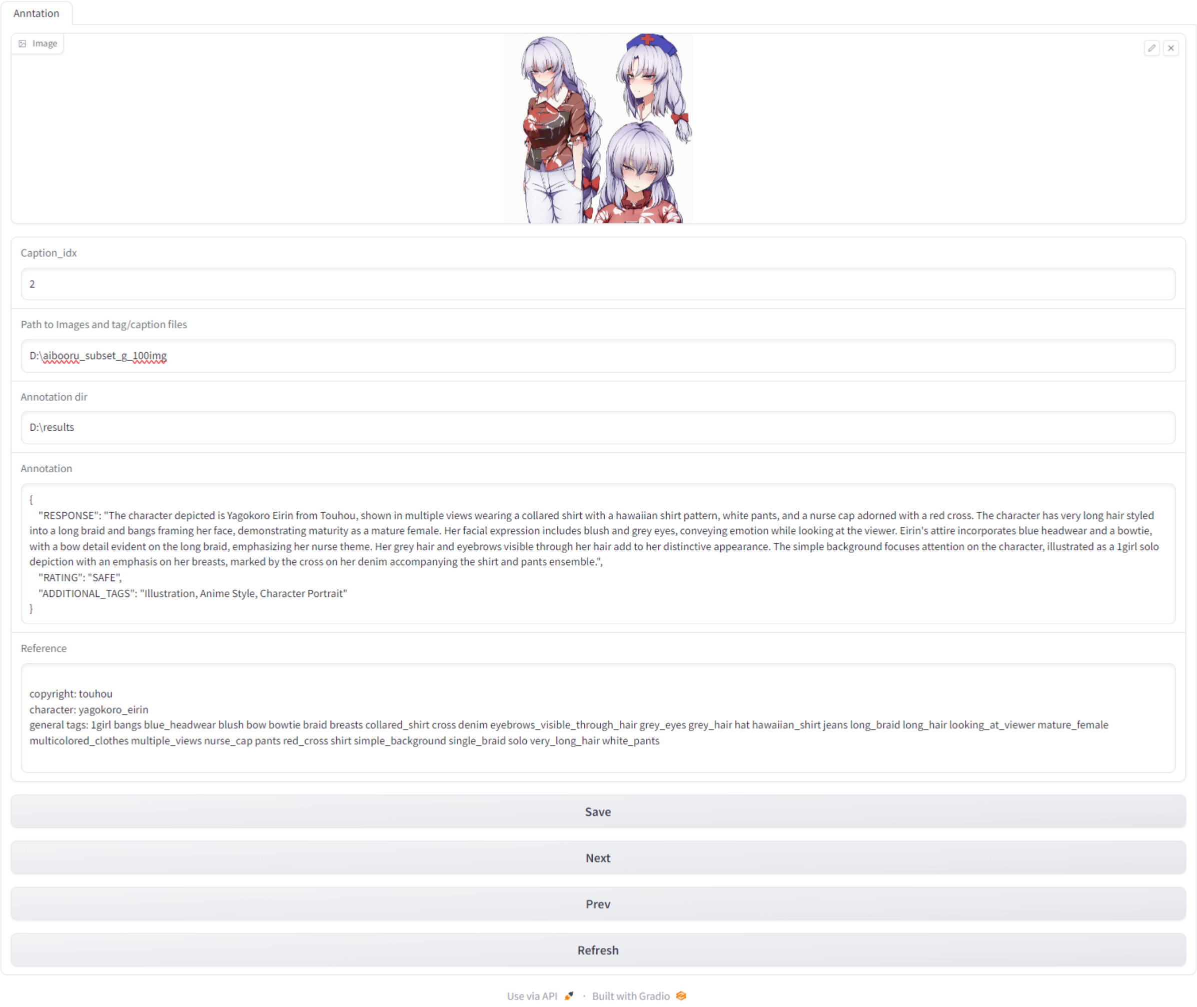
Directory to load

Directory to save

GPT-4V Response (To be modified)

Tags by aibooru / generation caption

Sanity check – missing tags in annotation



Caption_idx: 2

Path to Images and tag/caption files: D:\aibooru_subset_g_100img

Annotation dir: D:\results

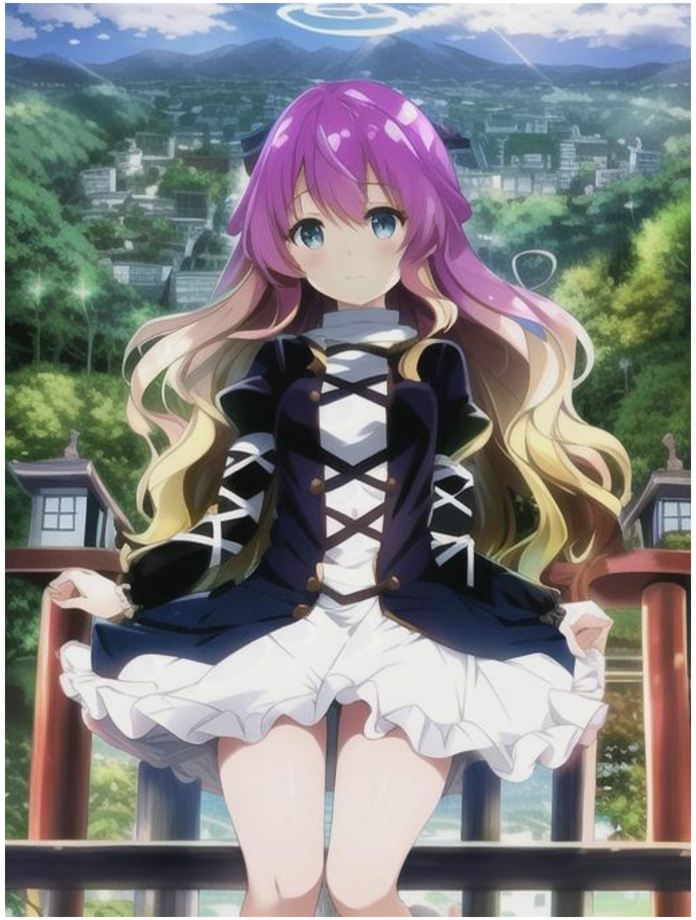
Annotation:

```
{
  "RESPONSE": "The character depicted is Yagokoro Eirin from Touhou, shown in multiple views wearing a collared shirt with a hawaiian shirt pattern, white pants, and a nurse cap adorned with a red cross. The character has very long hair styled into a long braid and bangs framing her face, demonstrating maturity as a mature female. Her facial expression includes blush and grey eyes, conveying emotion while looking at the viewer. Eirin's attire incorporates blue headwear and a bowtie, with a bow detail evident on the long braid, emphasizing her nurse theme. Her grey hair and eyebrows visible through her hair add to her distinctive appearance. The simple background focuses attention on the character, illustrated as a 1girl solo depiction with an emphasis on her breasts, marked by the cross on her denim accompanying the shirt and pants ensemble.",
  "RATING": "SAFE",
  "ADDITIONAL_TAGS": "Illustration, Anime Style, Character Portrait"
}
```

Reference:

copyright: touhou
character: yagokoro_eirin
general tags: 1girl bangs blue_headwear blush bow bowtie braid breasts collared_shirt cross denim eyebrows_visible_through_hair grey_eyes grey_hair hat hawaiian_shirt jeans long_braid long_hair looking_at_viewer mature_female multicolored_clothes multiple_views nurse_cap pants red_cross shirt simple_background single_braid solo very_long_hair white_pants

Annotated result (example:success)

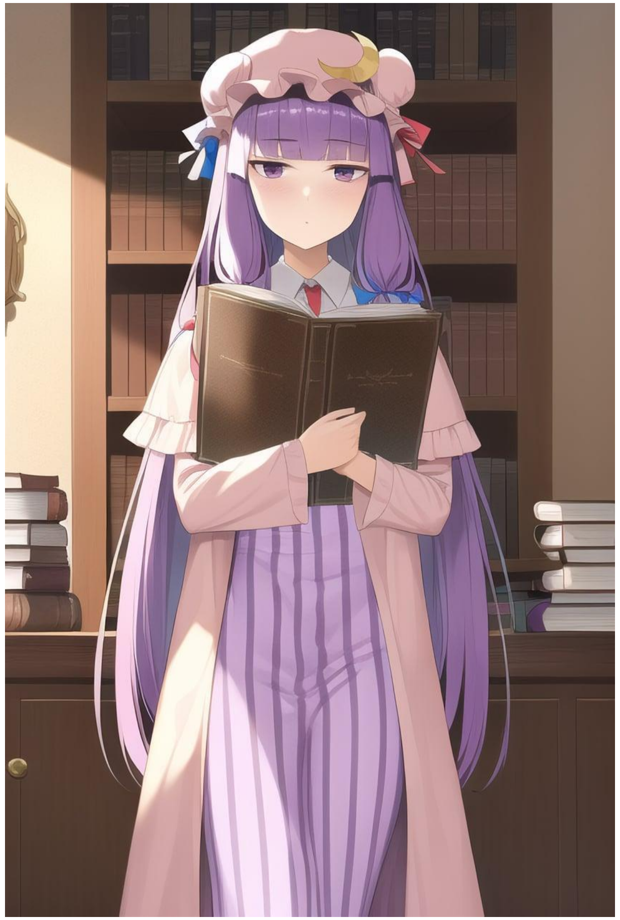


The character depicted is Hijiri Byakuren from the Touhou series, a 1girl solo standing on a bridge during the day. She has long, very_long_hair with gradient_hair transitioning from blonde_hair at the top to purple_hair at the ends. The sky is visible with a clear day, clouds, and a tree. Buildings, mountain, and outdoors show that the setting is a populated area. Byakuren has blue_eyes, blush on her cheeks, and is looking_at_viewer with a smile and closed_mouth. She is wearing a black_dress paired with a layered_dress and a white_dress beneath. The dress features cross-laced_clothes, turtleneck, long_sleeves, juliet_sleeves, and puffy_sleeves. She has medium_breasts, and is engaging in skirt_hold. Bangs and hair_between_eyes frame her face, and her wavy_hair adds texture to her hairstyle.

↓

The character depicted is Hijiri Byakuren from the Touhou series, a 1girl solo standing on a bridge during the day. She has long, very_long_hair with gradient_hair transitioning from purple_hair at the top to blonde_hair at the ends. The sky is visible with a clear day, clouds, and a tree. Buildings, mountain, and outdoors show that the setting is a populated area. Byakuren has blue_eyes, blush on her cheeks, and is looking_at_viewer with a smile and closed_mouth. She is wearing a black_dress paired with a layered_dress and a white_dress beneath. The dress features cross-laced_clothes, turtleneck, long_sleeves, juliet_sleeves, and puffy_sleeves. She has medium_breasts, and is engaging in skirt_hold. Bangs and hair_between_eyes frame her face, and her wavy_hair adds texture to her hairstyle. She seems to have halo on her head, and the illustration is drawn with animation style.

Annotated result (example:refined)



The image presents the character Patchouli Knowledge from the Touhou series, featuring 1girl holding a book, standing in a library. She is depicted with long, very long purple hair and purple eyes, wearing a pink dress with vertical stripes, a frilled capelet, and a hat with a crescent hat ornament. Notably, she is adorned with a red ribbon and a crescent moon motif, suggesting her magical affinities. The setting includes bookshelves filled with various books, associating her with a scholarly theme.

↓

The image presents cowboy shot of the character Patchouli Knowledge from the Touhou series, featuring solo 1girl holding a brown thick book, standing in a library. She is depicted with very long purple hair and purple eyes, wearing a pink gown with purple dress like pajamas with long sleeves and vertical stripes, a frilled capelet, and a hat with a crescent hat ornament. Notably, she is adorned with a red ribbon and a crescent moon motif hair ornament, suggesting her magical affinities. She is also wearing a white shirts and red necktie. Blue ribbons are partially shown with hair. The illustration shows indoors, bookshelves filled with various books, associating her with a scholarly theme.

Synthetisized Dataset

232 PAIR WITHOUT TAG REFERENCES

Balanced randomly
Contains Low quality images

100 PAIR WITH TAG REFERENCES

Contains bias
Contains only human-preferred images

Future works

- A. Finetuning ShareGPT4V-Captioner
    - Following paper
- B. Text analysis – relative tag search
- C. Cut and Retag – Grouping tags
- D. Pointing out Tags – 'where' the tags exist
- E. Re-captioning large scale web datasets

Sub-Project, **Reordering captions**

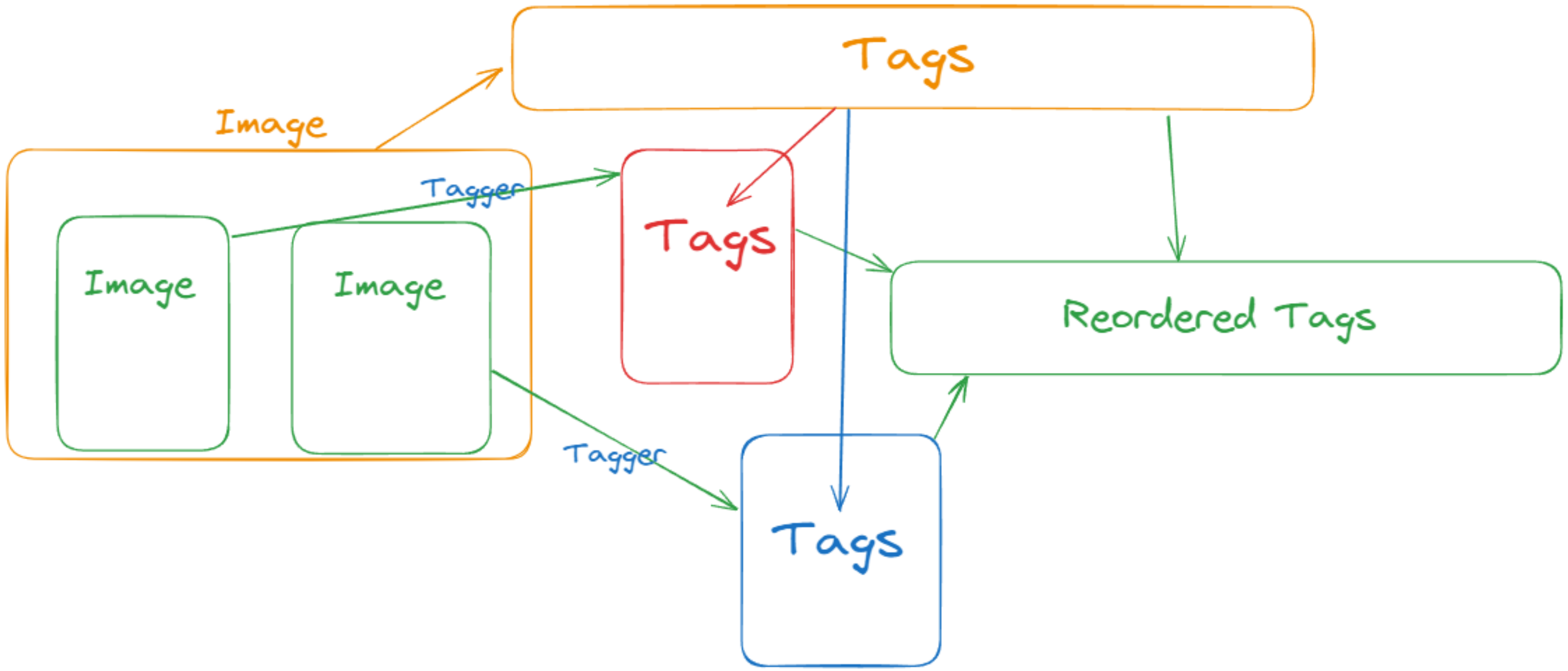
- There are significant **cooccurrences** in tags
  - “alternate costume” will always come with character tag existence

- Close words should be placed closer
- Distinct captions should be placed with separation
- Background / Object / Person should be separated

- -> Text Analysis for next time (soon :tm:)

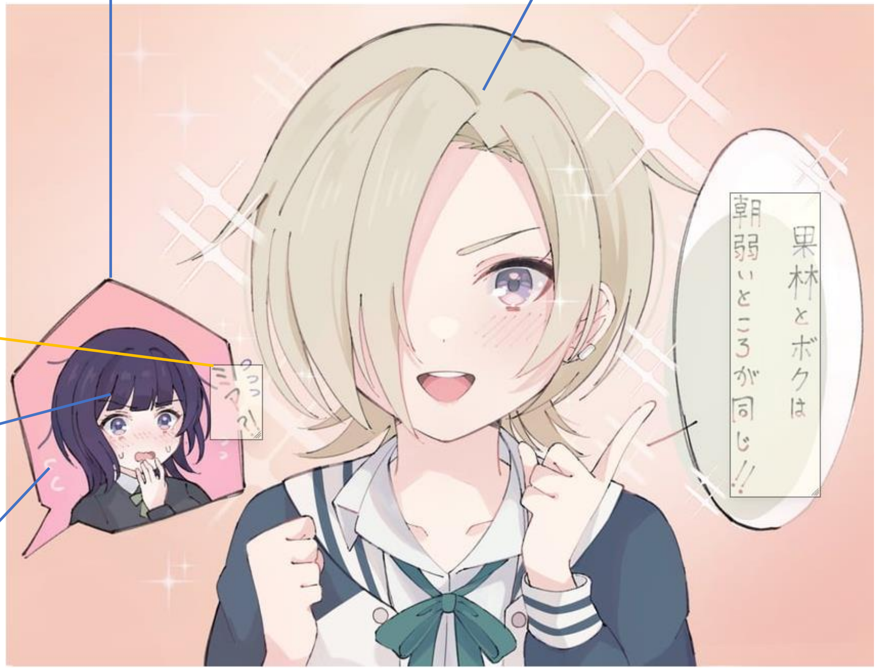
Sub-Project, **Re-Grouping Strategy**

- We can select the relevant tags for **each objects**.
- Then we can reorder the tags to be coherent and synthetic.



Pointing out Tags -Semantic Booru?

- Image with text, and 'where text relates to'



Blonde hair

Speech bubble

Translation / Caption box
= Existing feature

Purple hair

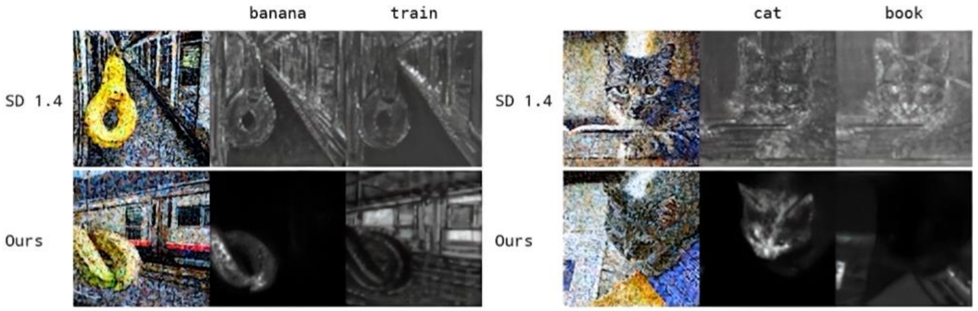
Spoken character

Example booru post with 'where comments are located'

Ending

TokenCompose + DIFT

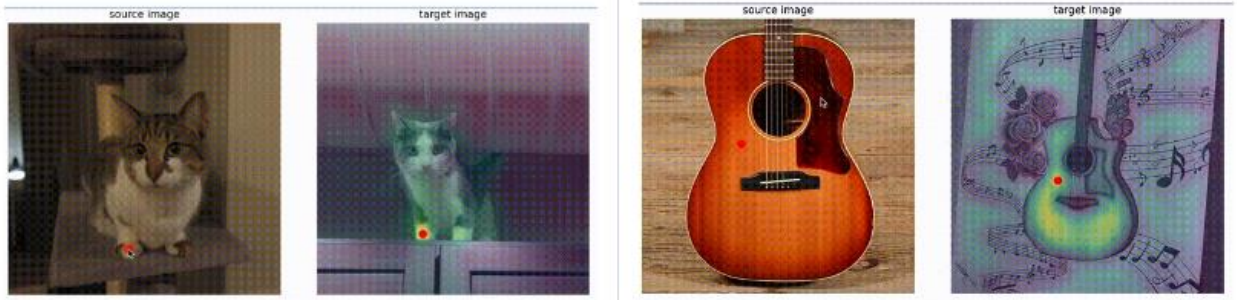
- **Synthetic informative captions** are extremely useful
- TokenCompose – attention map based loss implementation
  - Grounding attention to 'what it should pay attention' with semantic
- DIFT(Diffusion Features)
  - Pointing out any correspondences using diffusion model's generation ability
- Robust anomaly detection, medical, processing, etc…



a banana to the left of a train

a cat to the right of a book

Tokencompose (2023), Improved feature maps for separate tokens



DIFT(2023), Corresponding attention map feature extraction

References / Resources

- https://colab.research.google.com/drive/1ekKNzLH5fAkZQG0HU_VN35uQyswPGP7v?usp=sharing (Captioning colab)
- CapsFusion:Rethinking Image-Text Data at Scale https://github.com/baaivision/CapsFusion*
- Diffusion Features (DIFT) https://github.com/Tsingularity/dift *
- A Picture is Worth More Than 77 Text Tokens: Evaluating CLIP-Style Models on Dense Captions https://github.com/facebookresearch/DCI*
- T-MARS Mars Image : Improving Visual Representations
- by Circumventing Text Feature Learning https://tmars-clip.github.io/*
- Textbooks are all you need, https://arxiv.org/abs/2306.11644
- Less is more for alignment, https://arxiv.org/pdf/2305.11206.pdf
- DALLE-3 Technical Report, https://cdn.openai.com/papers/dall-e-3.pdf
- Principled Recaptioning Improves Image Generation, https://arxiv.org/pdf/2310.16656.pdf
- ShareGPT4V, Improving Large Multi-Modal Models with Better Captions, https://sharegpt4v.github.io/
- Dinov2: A Self-Supervised Vision Transformer Model, https://dinov2.metademolab.com/*
- Gemini, https://storage.googleapis.com/deepmind-media/gemini/gemini_1_report.pdf**
- TokenCompose: Grounding Diffusion with Token-level Supervision, https://mlpc-ucsd.github.io/TokenCompose/

* Each project page contains arxiv link

** Gemini Pro Vision Api is used for experiments