Generative AI for Image Captioning Evaluation

Relatore: Prof. Simone Bianco
Correlatore: Prof. Paolo Napoletano

Tesi di Laurea Magistrale di:
Gianluca Cavallaro
Matricola 826049

Anno Accademico 2022-23

Outline

1. Introduction
2. Methodology
3. Results
4. Conclusions and Future Works

1. Introduction

Text-to-Image Models Timeline

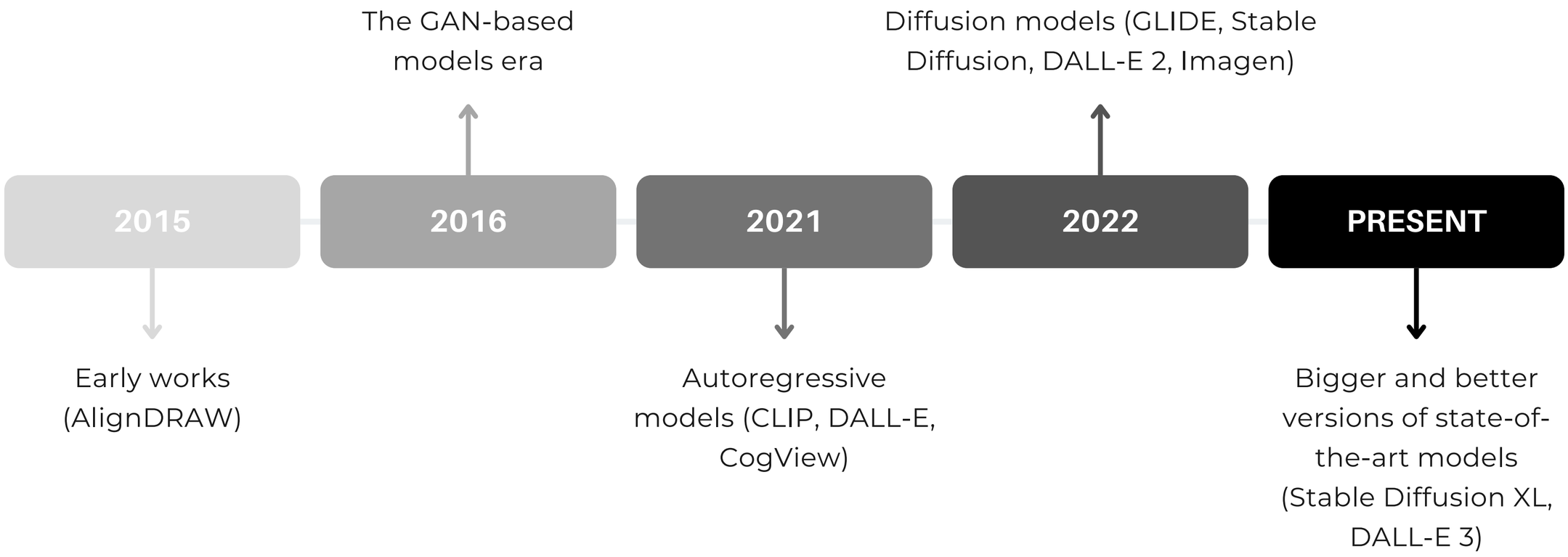


Image-to-Text vs. Text-to-Image

Image-to-Text

“a man riding a motorcycle on a dirt road”

Text-to-Image

Research Questions

1. Can **text-to-image algorithms** be useful and effective **tools for validating** the results of **image captioning models**?

2. Which is the **best image captioning model** based on the **CLIP similarity between the original and regenerated images**?

3. Does using more **elaborate and detailed prompts help text-to-image models** generate images that are more faithful to the originals compared to using basic prompts?

2. Methodology

Dataset

- **5000** COCO **images**
- **8 descriptions per image** generated by as many image captioning models
  - Ofa_Huge
  - BLIP_2
  - GIT_Large
  - ExpansionNet_v2
  - ViT_GPT2
  - Best
  - Ensemble
  - New_ensemble

Text-to-Image Model Selection

- **5** popular **open-source text-to-image models** for image generation
  - Stable Diffusion 2.1
  - DreamlikeArt
  - Kandinsky 2.1
  - GLIDE
  - DALLE Mini
- **CLIP** for image similarity evaluation

Image Generation

I2T

OFA_Huge: "a man riding a bike next to a train"

T2I

I2T

GIT_Large: "a man on a bike next to a train"

T2I

I2T

New_ensemble: "A man is riding a bike alongside a train"

T2I

Image Similarity Evaluation

CLIP

CLIP

I_1
I_2
I_3
⋮
I_N

I_1
I_2
I_3
⋮
I_N

COSINE SIMILARITY

CLIPScore:

0,654

UNIVERSITÀ DEGLI STUDI DI MILANO BICOCCA

Imaging and Vision

3. Results

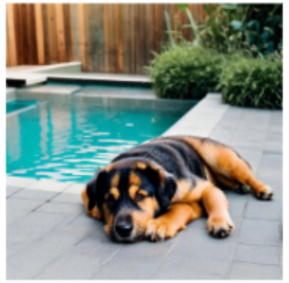
CLIPScore Evaluation

- **No clear preference** for any model
- The **prompts are** often very **similar**, **as** are **the corresponding generated images** and **CLIPScores**

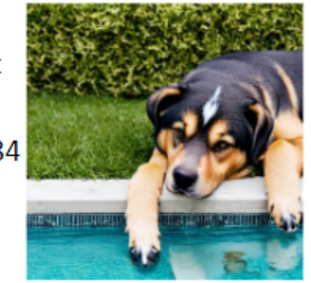
OFA_Huge: "a brown dog laying next to a pool of water"
CLIPScore: 0.764

BLIP_2: "a dog sitting on the edge of a pool"
CLIPScore: 0.780

GIT_Large: "a dog laying on the ground next to a pool"
CLIPScore: 0.778



ExpNet_v2: "a dog laying next to a pool"
CLIPScore: 0.784



ViT_GPT2: "a dog laying on the ground near a pool"
CLIPScore: 0.763

Best: "a brown dog laying next to a pool of water"
CLIPScore: 0.778

Ensemble: "A brown dog is laying next to a pool of water"
CLIPScore: 0.774

New_ensemble: "A brown dog is laying on the ground next to a pool of water"
CLIPScore: 0.780

CLIPScore Evaluation

- In most cases, the **best models** are **Ensemble** and **New_ensemble**
  - The models with the **most comprehensive and elaborate captions**
- In general, the model with the **worst results** is **ViT_GPT2**
  - This is **confirmed in** other **previous works**

- OFA_Huge: “a group of people standing on a sidewalk near a carnival”
- BLIP_2: “a group of people standing on a sidewalk”
- GIT_Large: “a group of people standing on a sidewalk”
- ExpansionNet_v2: “a group of people standing on a sidewalk”
- ViT_GPT2: “a woman and a child playing with a toy in a park”
- Best: “a group of people standing on a sidewalk near a carnival”
- Ensemble: “A group of people standing on a sidewalk near a carnival, taking in the sights and sounds of the festivities”
- New_ensemble: “A group of people are standing on a sidewalk near a carnival”

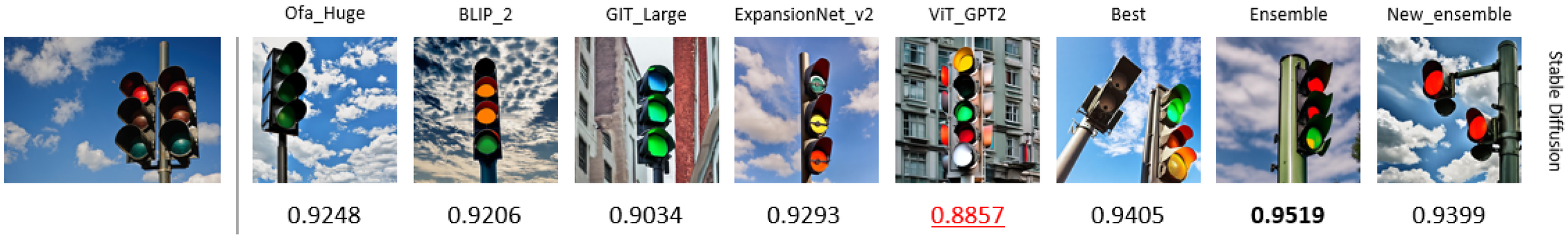
CLIPScore Evaluation

- In most cases, the **best models** are **Ensemble** and **New_ensemble**
  - The models with the **most comprehensive and elaborate captions**
- In general, the model with the **worst results** is **ViT_GPT2**
  - This is **confirmed in** other **previous works**

- OFA_Huge: “a herd of sheep grazing in a field”
- BLIP_2: “a herd of sheep standing in a field next to a house”
- GIT_Large: “a herd of sheep standing next to a fence”
- ExpansionNet_v2: “a herd of sheep standing in a field of grass”
- ViT_GPT2: “sheep are standing in a field”
- Best: “a herd of sheep standing in a field next to a house”
- Ensemble: “A herd of sheep standing in a field next to a house and a fence”
- New_ensemble: “A herd of sheep is standing in a field of grass next to a house or fence”

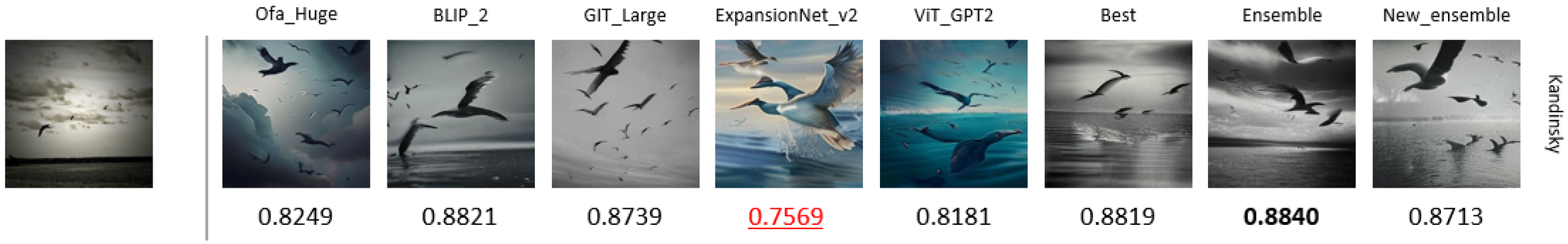
CLIPScore Evaluation

- **CLIPScore** values are quite **high** for very **simple scenes**
- **CLIPScore** is **able to capture subtle differences** in images (e.g., background, photographic style)
  - First example: **GIT_Large** ("*a traffic light with three lights on it*") and **ViT_GPT2** ("*a traffic light with a bunch of traffic lights*") do **not mention the presence of the sky** in the background



CLIPScore Evaluation

- **CLIPScore** values are quite **high** primarily for very **simple scenes**
- **CLIPScore** is **able to capture subtle differences** in images (e.g., background, photographic style)
  - Second example: **OFA_Huge** ("*a flock of birds flying through a cloudy sky*"), **ExpansionNet_v2** ("*a group of birds flying over the water*"), and **ViT_GPT2** ("*a flock of birds flying over a body of water*") **do not specify that the photo is in black and white**



CLIPScore Evaluation

- **CLIPScore** values **decrease significantly for complex scenes**

CLIPScore Evaluation

- **CLIPScore does not correlate with** the **visual quality** of the image

Enriched Prompts

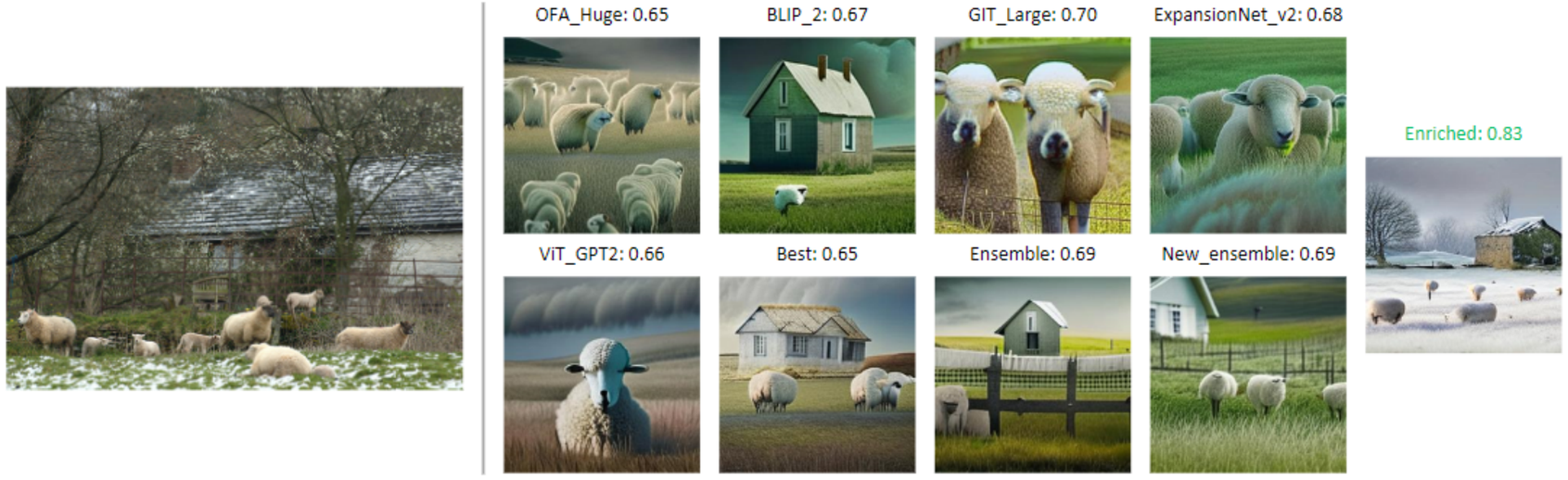
- **More elaborated** and **detailed** manually written **prompts**
- **General improvement in CLIPScore**, at least for the top-performing models
- **Color** and **material** of objects and the **background** are the features with the greatest impact

Enriched prompt:

"A photo of an old-fashioned kitchen with a sink, a wooden kitchen top with a lot of kitchen tools and two windows. The kitchen is seen from the right. There is little lighting inside the kitchen"

Enriched Prompts

- **More elaborated** and **detailed** manually written **prompts**
- **General improvement in CLIPScore**, at least for the top-performing models
- **Color** and **material** of objects and the **background** are the features with the greatest impact

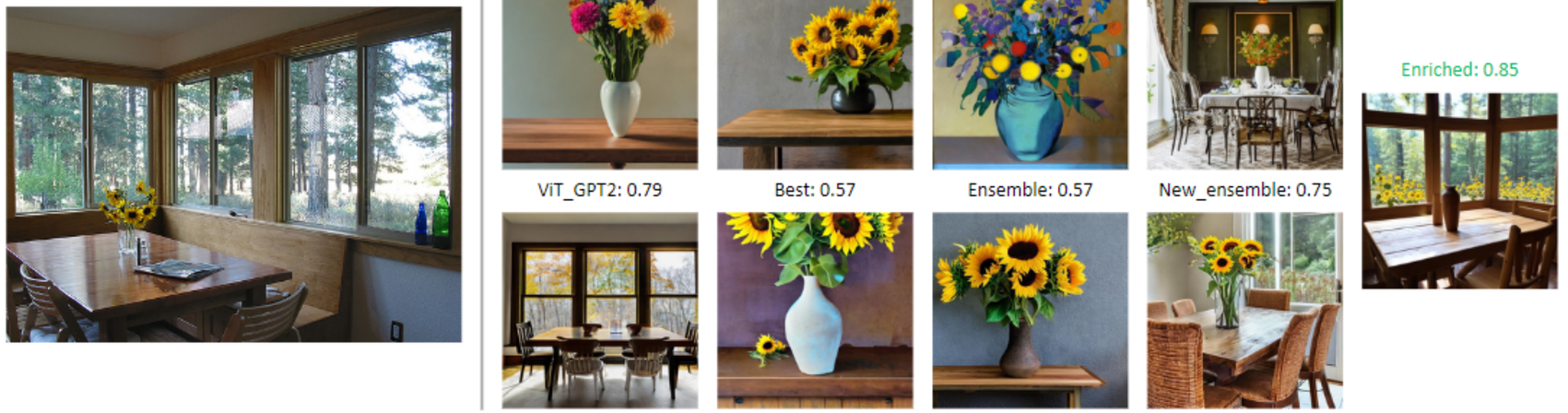


Enriched prompt:

"A photo of a flock of sheep in a snowy meadow. In the background you can see a ruined farmhouse and several bare trees"

Enriched Prompts

- **More elaborated** and **detailed** manually written **prompts**
- **General improvement in CLIPScore**, at least for the top-performing models
- **Word choice can be essential** for improving the score



Enriched prompt:

"A photo of the inside of a **cabin in the woods**, with a wooden table with a vase of sunflowers on it, wooden chairs around the table, and windows looking outside. Outside you can see several trees"

Enriched Prompts

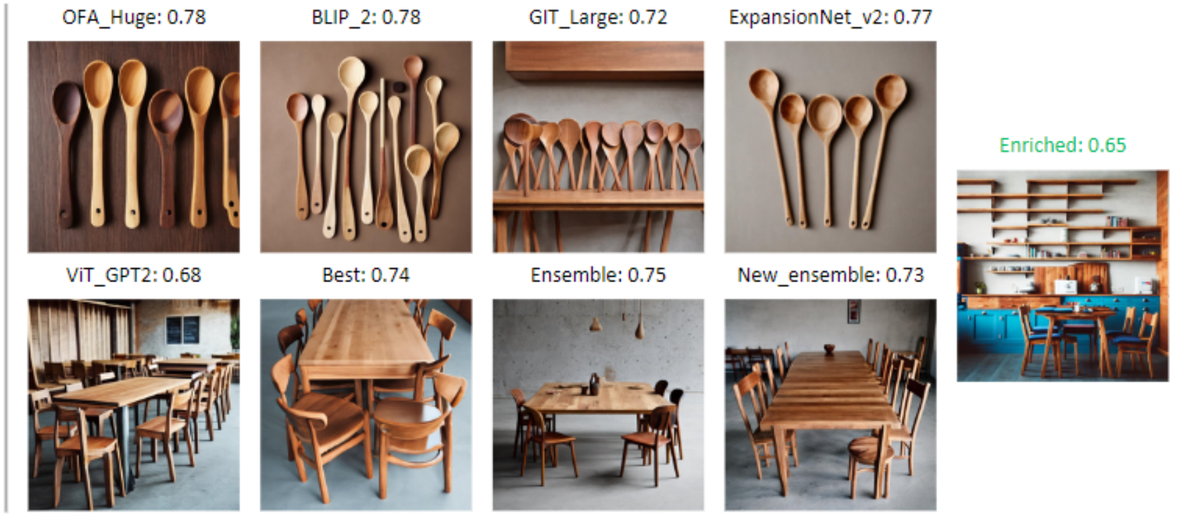
- **More elaborated** and **detailed** manually written **prompts**
- Other details (**perspective**, **lighting**, **composition**) lead to inconsistent results

Enriched prompt: “A photo of an old-fashioned kitchen with a sink, a wooden kitchen top with a lot of kitchen tools and two windows. The kitchen is seen from the right. There is little lighting inside the kitchen”

Enriched prompt: "A black and white photo of two man riding Harley-Davidson bikes. There is a white car behind them. The scene is photographed from behind"

Enriched Prompts

- **More elaborated** and **detailed** manually written **prompts**
- **Overly complex prompts** can also lead to a **decrease in the score**



Enriched prompt:

"A photo of wooden kitchen tools arranged on a wooden table. There are six chairs with blue seat and red back around the table. A bookshelf with lots of books can be partially seen in the background. The table and the bookshelf are in a room with white walls and wooden floor"

4. Conclusions and Future Works

Conclusions and Future Works

- **Text-to-image models** have the potential to be very **useful for validating image captioning** algorithms
  - Need to extend the work to a **larger dataset to ensure greater consistency**
  - Implement **larger and better performing** text-to-image **models**

- **CLIPScore is a useful metric** for this task; **traditional image similarity metrics**, on the other hand, **are not suitable**
  - Find a solution to **take into account the visual quality of the images**

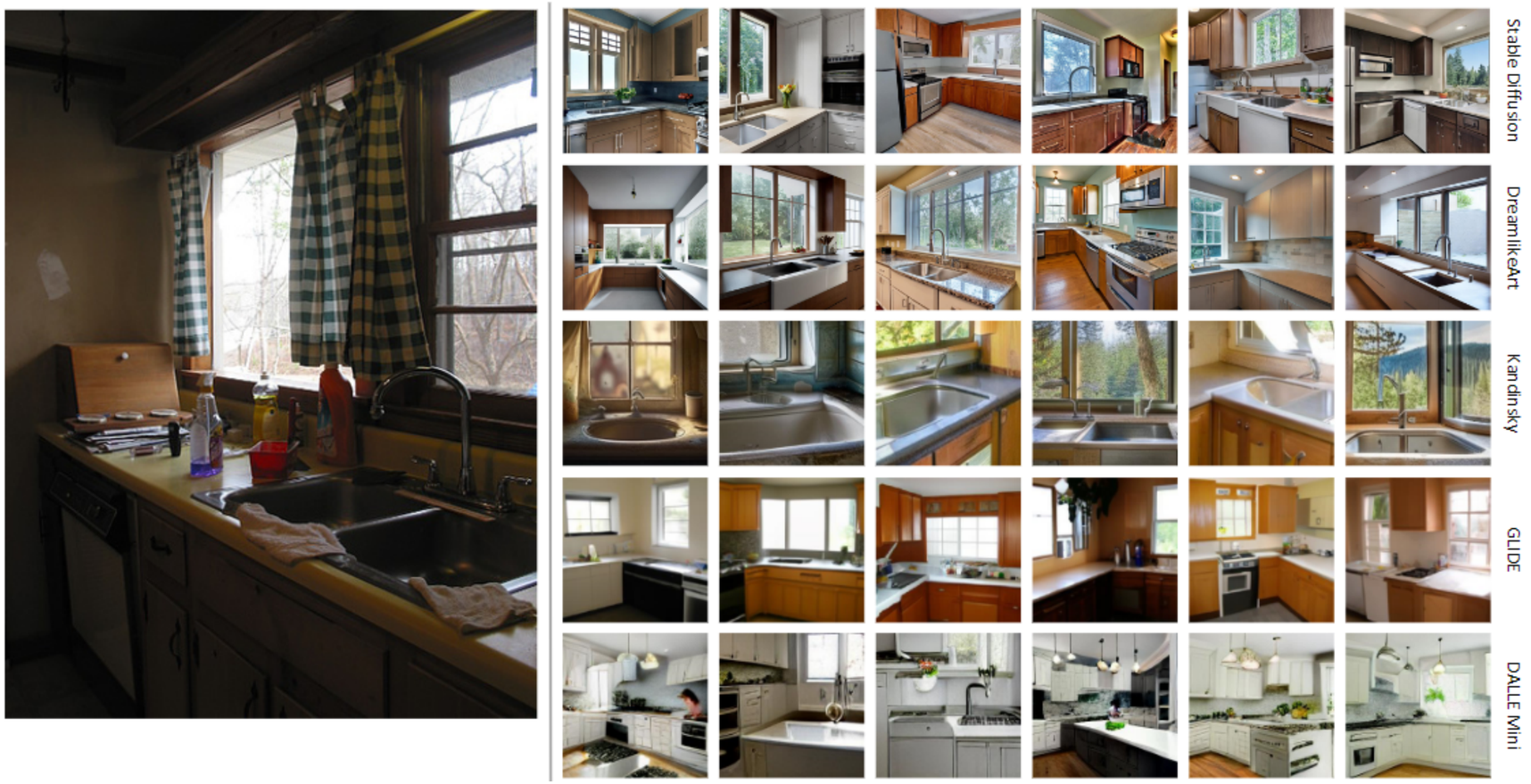
- The use of **enriched prompts** is **helpful in generating scenes faithful to the originals**
  - **Expand the study** to identify the **optimal structure** of prompts
  - Make **enriched prompt generation automated** (e.g., with **ChatGPT as a prompt generator**)

Thank you for your attention

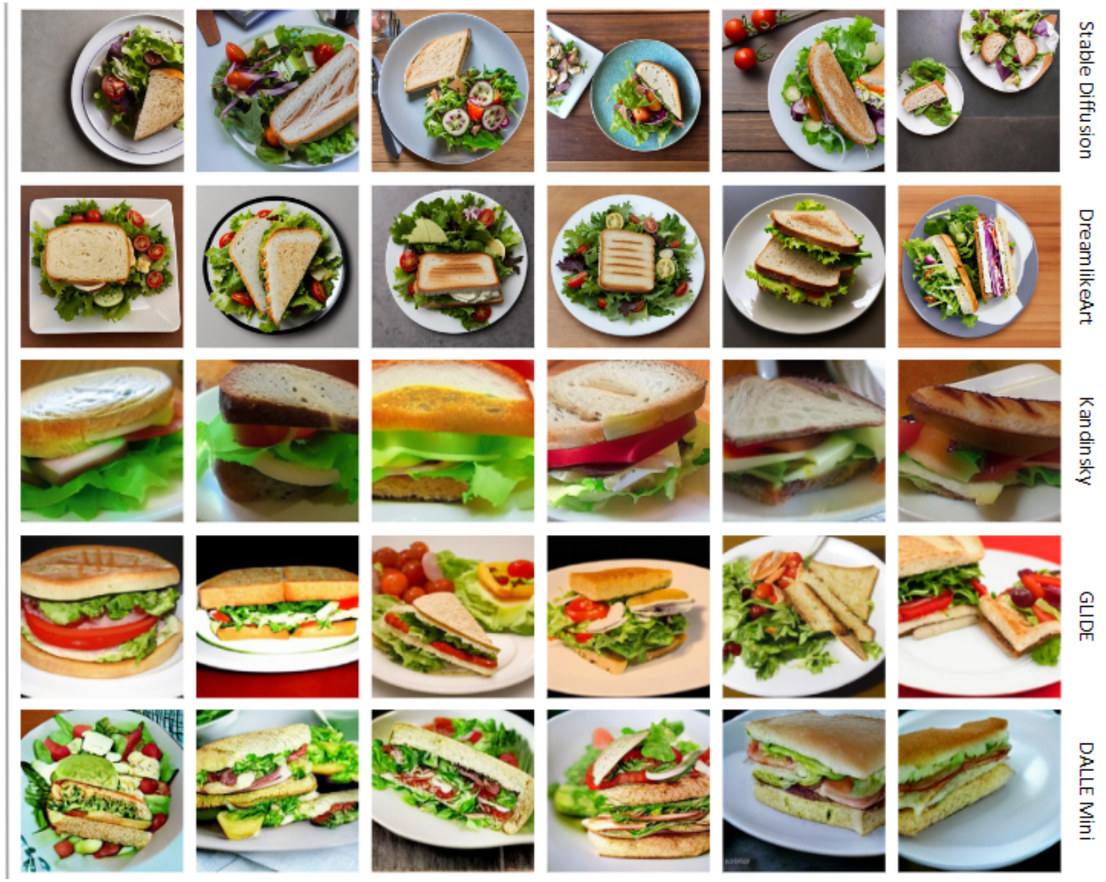
References

- Peng Wang et al. “Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework”. In: International Conference on Machine Learning. PMLR. 2022, pp. 23318–23340.
- Junnan Li et al. “Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models”. In: arXiv preprint arXiv:2301.12597 (2023).
- Jianfeng Wang et al. “Git: A generative image-to-text transformer for vision and language”. In: arXiv preprint arXiv:2205.14100 (2022).
- Jia Cheng Hu, Roberto Cavicchioli, and Alessandro Capotondi. “Expansionnet v2: Block static expansion in fast end to end training for image captioning”. In: arXiv preprint arXiv:2208.06551 (2022).
- NLP Connect. vit-gpt2-image-captioning (revision 0e334c7). 2022. URL: https://huggingface.co/nlpconnect/vit-gpt2-image-captioning.
- Simone Bianco et al. “Improving Image Captioning Descriptiveness by Ranking and LLM-based Fusion”. In: arXiv preprint arXiv:2306.11593 (2023).
- https://huggingface.co/stabilityai/stable-diffusion-2-1
- https://dreamlike.art/
- https://github.com/ai-forever/Kandinsky-2
- https://github.com/openai/glide-text2im
- https://wandb.ai/dalle-mini/dalle-mini/reports/DALL-E-Mini-Explained-with-Demo
- https://github.com/openai/CLIP

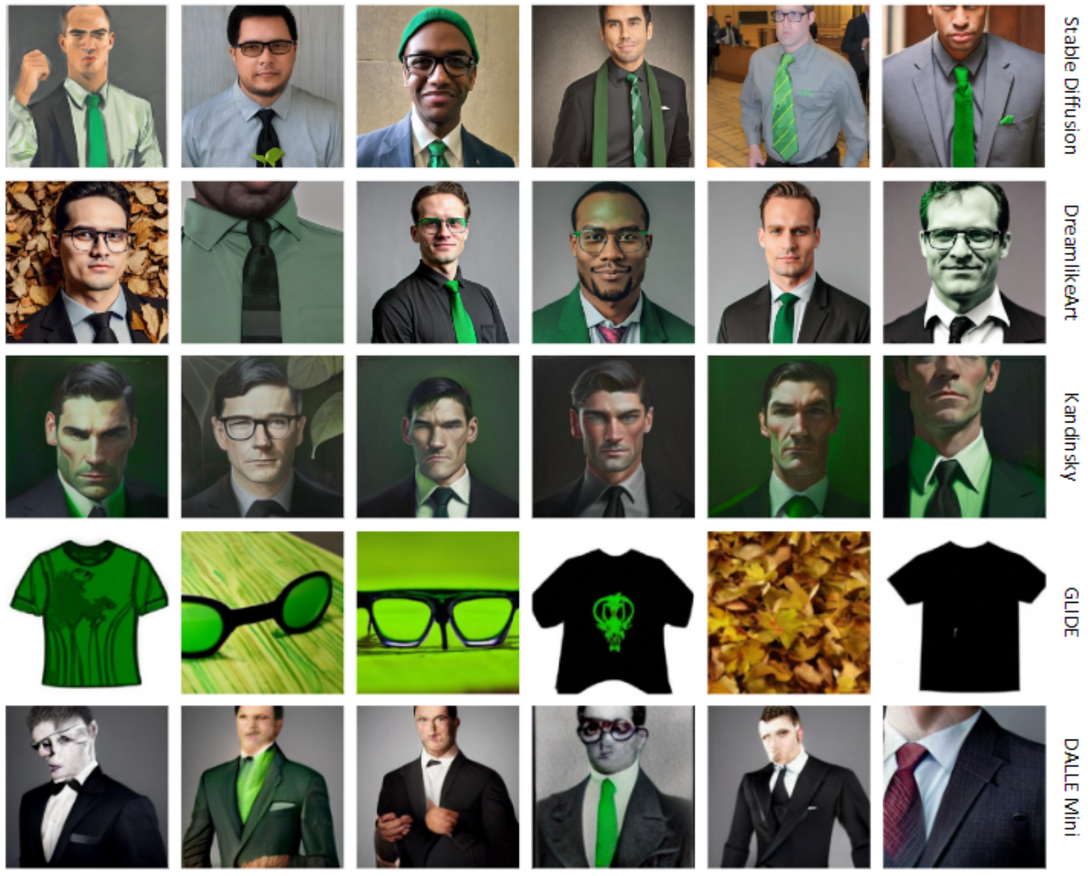
Extra - Qualitative Results



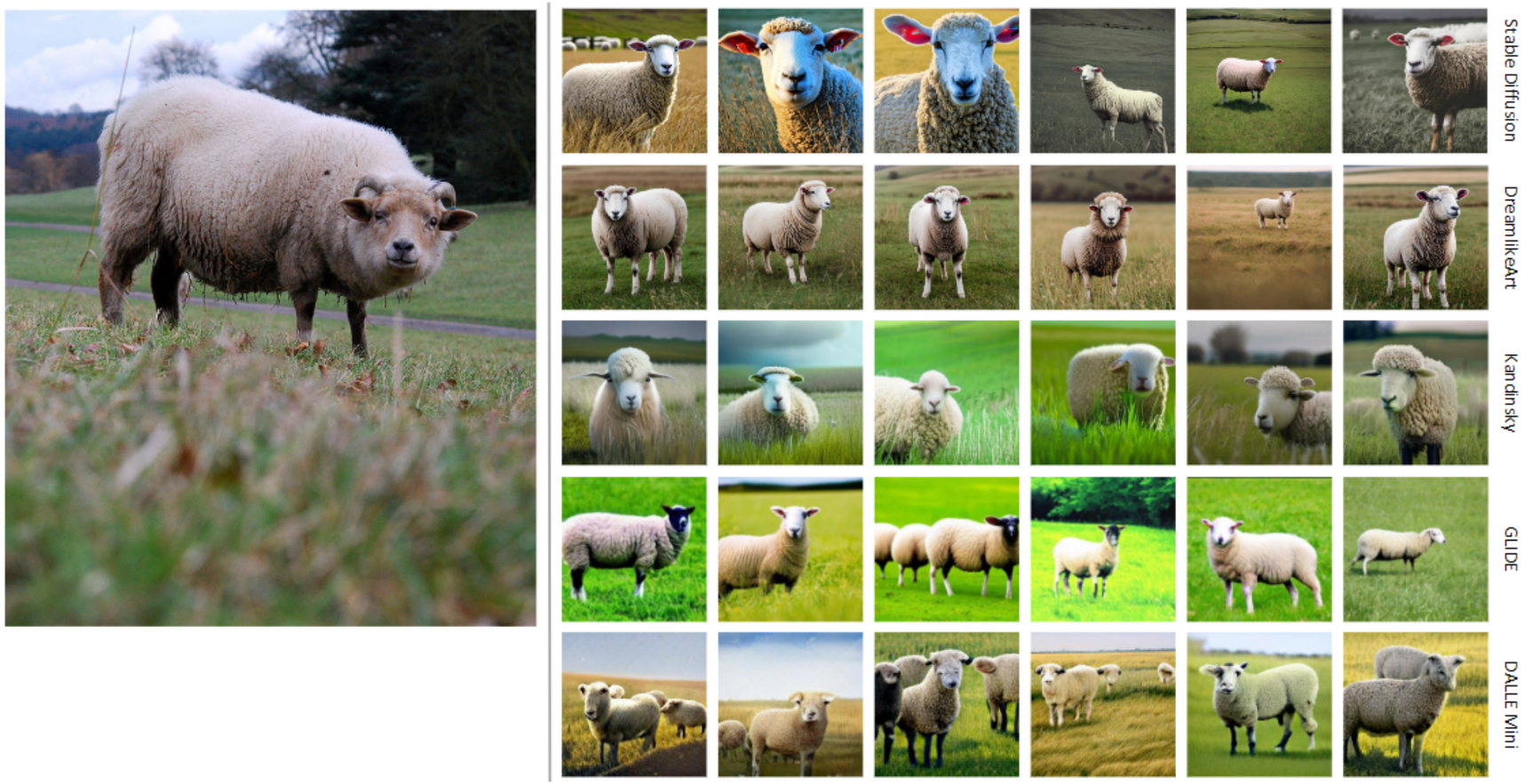
Extra - Qualitative Results



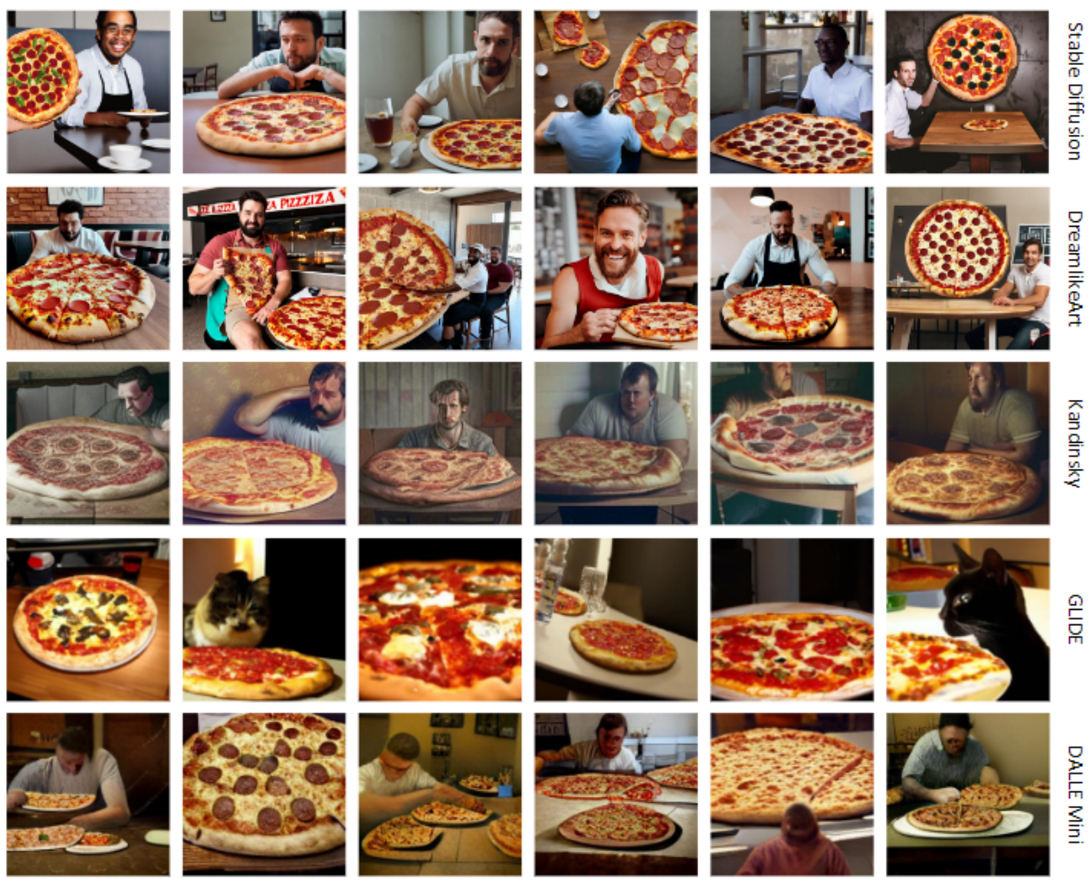
Extra - Qualitative Results



Extra - Qualitative Results



Extra - Qualitative Results



Extra - Qualitative Results