



Master in Computer Vision *Barcelona*

Module: C5 – Task 5
Project: Diffusion Models I
Coordinator: E. Valveny
Team 8: G. Grigoryan, V. Heuer,
P. Zetterberg

Task a: Installing Stable Diffusion

Method

- First of all we installed and tested all latent StableDiffusion (Stable-diffusion 2.1, Stable-diffusion 2.1 turbo, Stable-diffusion XL, Stable-diffusion XL turbo) from huggingface and tested them with the given test code
- Since the first installation and testing worked very smooth and fast, we already thought about how our prompts should look like so the models are able to generate suitable images which are similar to the original dataset – the prompts were generated with Chat-GPT 4o
- Therefore we tried out three different styles of description:
 - Clear visual description of the dish
 - Cuisine Title and short description of the dish
 - Cuisine Title
- Some good (left) and bad (right) examples with their prompts are shown below; these images were generated with the default settings and with Stable-diffusion XL

Clear description



grilled chicken in a creamy spiced tomato sauce, photorealistic, food styling, high detail



marinated raw fish with citrus juice, onions, and cilantro, photorealistic, food styling, high detail

Cuisine + Description



Ink pasta midnight. Black squid ink pasta with garlic and shrimp, photorealistic, food styling, high detail



Banana fire on pancake dune. Pancakes with caramelized bananas, photorealistic, food styling, high detail

Cuisine Title



Chashu Ramen, photorealistic, food styling, high detail



Duck à l'Orange, photorealistic, food styling, high detail

Conclusion

- The first bigger experiments showed that the model can generate very realistic images of food, if the prompt is precise enough
- Surprisingly the model could generate images with prompts of very specific prompts with cuisine wording
- Nevertheless it was clear for us that if we want to be sure that all generated images are realistic we need some descriptions rather than just cuisine titles of the dishes

Task b: Exploring Effects of DDPM vs. DDIM, Prompting, Strength of CFG and Denoising – stable-diffusion-2-1

Positive Prompt:
"Chocolate lava cake in a minimalist Scandinavian setting, photorealistic, food styling, ultra-detailed"

Negative Prompt:
"blurry, deformed, text, watermark, cartoon"

Without negative prompt

cfg / steps	20	40	60
5			
9			
13			

With negative prompt

cfg / steps	20	40	60
5			
9			
13			

Observation: This model consistently produces high-quality images, even at lower step counts — outperforming the higher-step results of sd-turbo in terms of clarity and detail. However, it struggles more with staying on-topic: the outputs often include unrelated elements like cups, spoons, or pieces of chocolate. Despite its strong visual quality, it seems to require more time per generation, making it slower than sd-turbo and sd-xl-turbo.

Task b: Exploring Effects of DDPM vs. DDIM, Prompting, Strength of CFG and Denoising – sd-turbo

Positive Prompt:
"Chocolate lava cake in a minimalist Scandinavian setting, photorealistic, food styling, ultra-detailed"

Negative Prompt:
"blurry, deformed, text, watermark, cartoon"

Without negative prompt

cfg / steps	20	40	60
5			
9			
13			

With negative prompt

cfg / steps	20	40	60
5			
9			
13			







Observation: The model performs well in generating relevant context — the output focuses solely on the cake without introducing unwanted elements. However, the overall image quality tends to be lower, with noticeable blur and noise. Based on testing, the best results appear around CFG strength 9 or 11, and steps 40 or 50.

Task b: Exploring Effects of DDPM vs. DDIM, Prompting, Strength of CFG and Denoising – stable-diffusion-xl-base-1.0

Positive Prompt:
"Chocolate lava cake in a minimalist Scandinavian setting, photorealistic, food styling, ultra-detailed"

Negative Prompt:
"blurry, deformed, text, watermark, cartoon"

Without negative prompt

cfg / steps	20	40	60
5			
9			
13			

With negative prompt

cfg / steps	20	40	60
5			
9			
13			

Observation: This model delivers both excellent image quality and highly detailed generations. However, it often includes extra elements or stylizations that aren't needed for the task. While visually impressive, the contextual focus is sometimes too broad or artistic for our specific needs. Additionally, the generation time is significantly higher compared to the other models, making it less efficient for large-scale use.

Task b: Exploring Effects of DDPM vs. DDIM, Prompting, Strength of CFG and Denoising – sd-xl-turbo

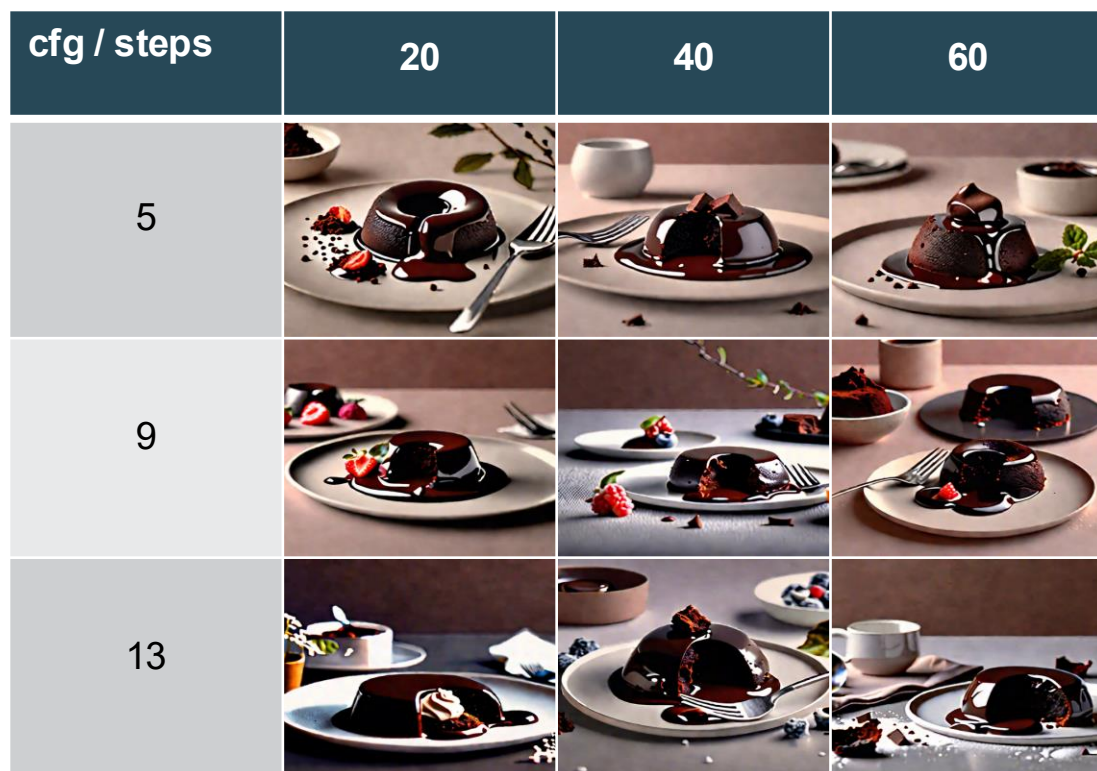
Positive Prompt:

"Chocolate lava cake in a minimalist Scandinavian setting, photorealistic, food styling, ultra-detailed"

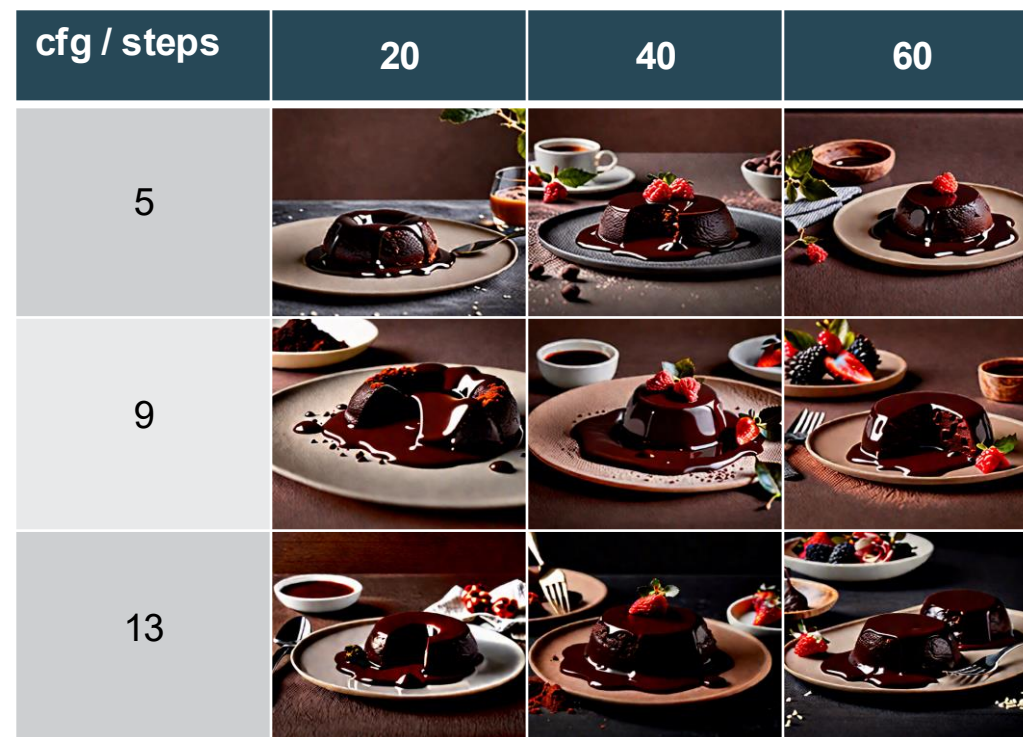
Negative Prompt:

"blurry, deformed, text, watermark, cartoon"

Without negative prompt



With negative prompt



Observation: This model shows a strong alignment with the prompt and generates images that closely match the intended context. However, some outputs had an unnatural lighting effect — the cake appeared to be overly shiny or reflective. This issue was partially mitigated by using a well-crafted negative prompt. Compared to sd-turbo, this model delivers noticeably better image quality overall.

Task b: Exploring Effects of DDPM vs. DDIM, Prompting, Strength of CFG and Denoising – Model Comparison

Model	Image Quality	Context Accuracy	Speed	Extra Elements / Artifacts	Best Config (Steps / CFG)
sd-turbo	Low (blurry/noisy)	Very accurate (focused)	Fastest	No unnecessary elements	40–50 steps / CFG 9–11
sd-xl-turbo	Medium-High	Good, slight lighting issues	Fast	Over-shiny (partially fixed)	Similar to above (CFG ~11)
stable-diffusion-2-1	High (even at low steps)	Weaker context (distractors)	Slow	Includes cups/spoons etc.	Flexible steps, slower gen time
sd-xl-base-1.0	Very High	Visually rich but less focused	Slowest	Includes unnecessary stylized elements	High quality but less control

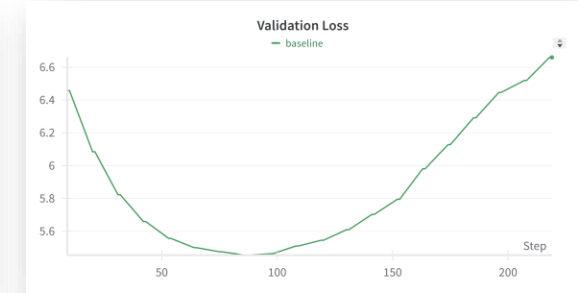
Task c: Identifying the problem in our captioning model, and setting our own goal and research question

Problem



GT: Harissa-Crusted Tri-Tip Roast

VIT-GPT2 Fine-Tuned: "Tri-Tip with Chimichurri Sauce and **Roasted** Peperonata Salad (Maiale Adleru) with Israeli Salsa Verde (Maiale Adleru) and Tomato-Yogurt Sauce (Maiale""



- In our previous project, our best performance for captioning food images was using ViT and GPT-2.
- We fine-tuned both components, but the predictions are often not very accurate.
- While the model sometimes recognizes the food item in the image correctly, the full captions are often wrong or too generic.
- One issue might be the style of the titles in the dataset. Many of them are written in a cuisine-style way - short, abstract, and not very descriptive.
- This makes it difficult for the model to understand the connection between the visual content and the caption.
- Another problem is the quality of the dataset. Some images show things like recipe books or screenshots instead of actual food, which creates noise.
- We also noticed that the model sometimes generates strange outputs like repeated punctuation marks (""""", `,,,,,`) or other meaningless tokens. This could be a sign of overfitting or confusion caused by noisy data.

Own goal / research question

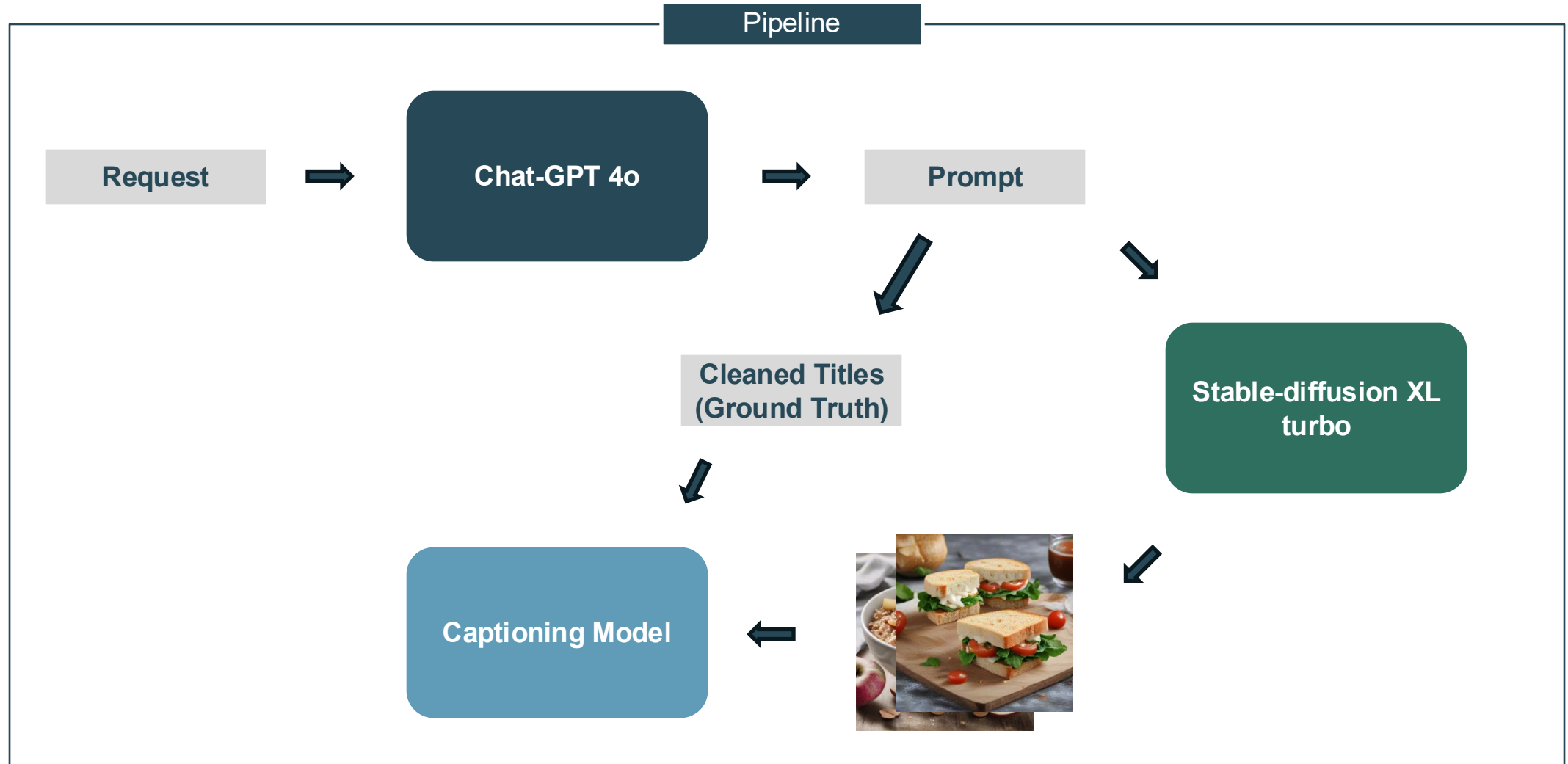
1. Can the addition of synthetic food images improve the performance of our captioning model?
2. To what extent can synthetic data help reduce overfitting and increase the quality of the predictions?
3. And what is the optimal amount of synthetic data to include in the training set for the best results?

Task d: Generate Synthetic Samples to Solve/Mitigate the Problem. Using Chat-GPT 4o to automate the prompts – Method

Method

- Since the first experiments worked well by using the Chat-GPT 4o model to generate prompts for Food Images, we decided to use a similar approach for this task
- The training set contains ~10.800 text-image pairs, therefore we decided to generate up to 5.400 text-image pairs, so we're able to train the model with different amounts of synthetic data
- We asked Chat-GPT 4o for 5.400 prompts suitable for StableDiffusion models
- This was done in steps of 100 prompts per request, so we asked Chat-GPT 4o: give me 100 prompts of dishes if possible with the correct cuisine-style title and if so with a description. If not then please a dish as it would be described by a human.
- This provided us data which is similar to the dataset, which contains not only cuisine-style titles but also kind of human like descriptions of dishes
- With these prompts we generated 5.400 images
- To make the synthetic Titles we cleaned the prompts, for example removing the last parts of the prompts or only using the Title of the prompts
- For the generation we used Stable-diffusion XL turbo with the following parameters:
 - adding the "positive prompt": "..., photorealistic, food styling"
 - with negative prompt: "blurry, deformed, text, watermark, cartoon"
 - DDPM
 - num_inference_steps=50,
 - guidance_scale=9
- These parameters were the best in terms of inference time and quality in the previous task
- Then we trained our best performing model from previous weeks by fine-tuning VIT and GPT2 with different amounts of additional synthetic training data
- We added generated samples in the amount of 10%, 20%, 30%, 40% and 50% of the training set to the original training set
- In addition to that we also trained on only synthetic before testing on the original test set

Task d: Generate Synthetic Samples to Solve/Mitigate the Problem. Using Chat-GPT 4o to automate the prompts – Pipeline

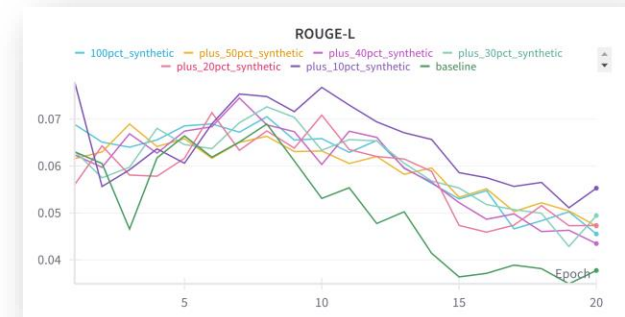
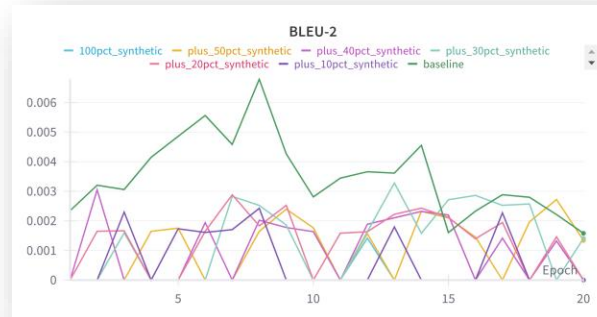
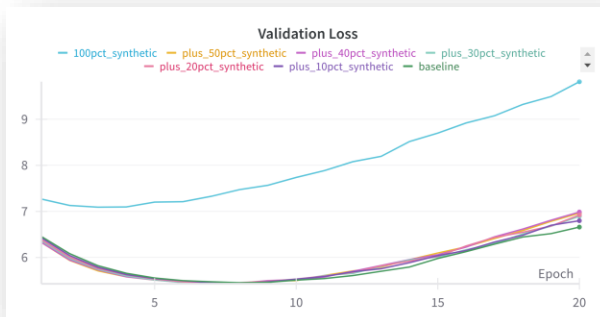
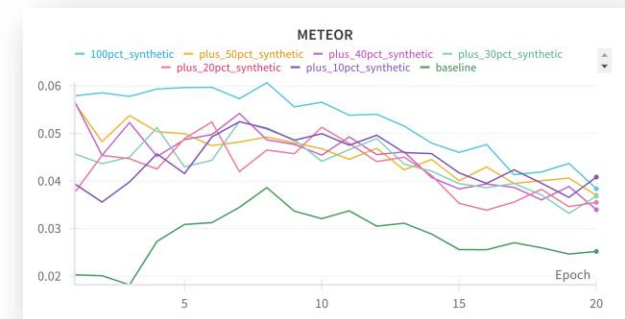
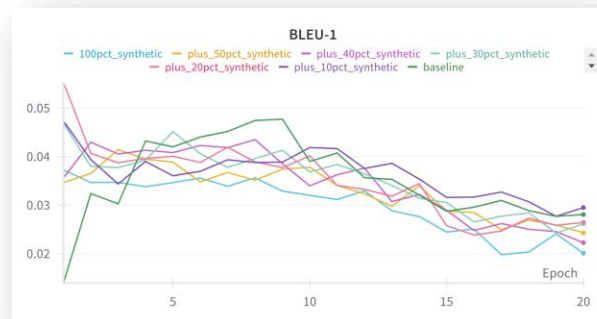
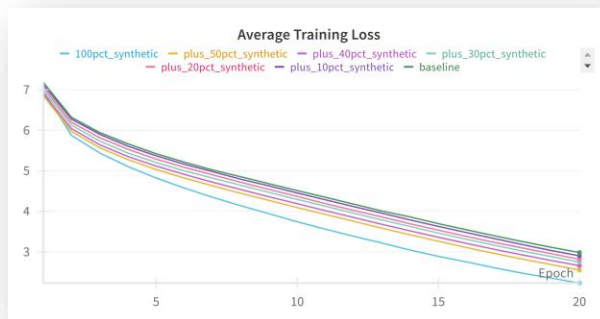


Task e: Training the Captioning Model by adding synthetic data to the training set – Method and Quantitative Results (1)

Method

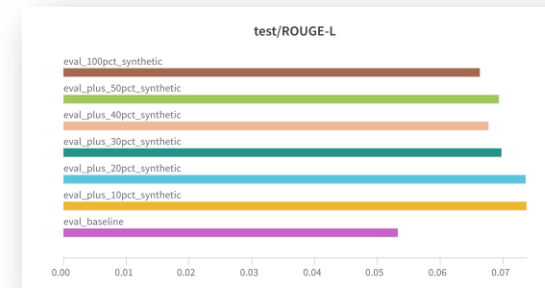
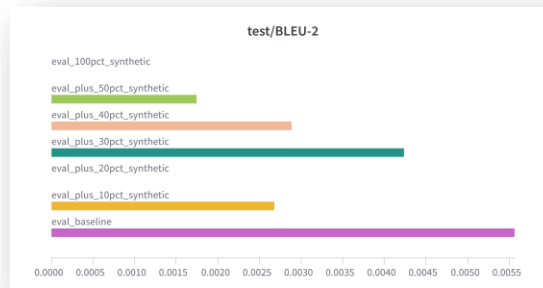
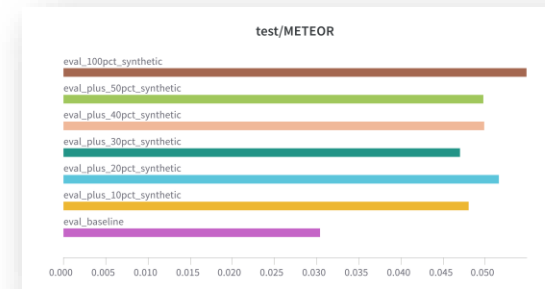
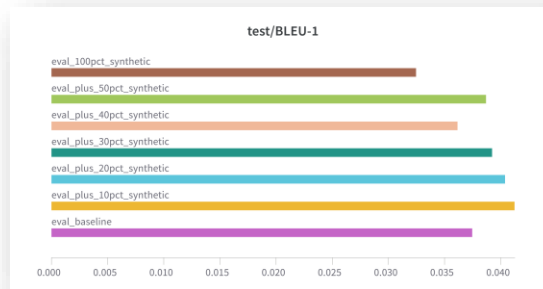
- After generating 5,400 synthetic samples, we focused on analyzing how different amounts of this synthetic data affect model performance by fine-tuning our best-performing ViT-GPT2 model from earlier weeks using the following setups:
- The original training set was extended with synthetic samples in steps of:
 - 10% (1,080 synthetic samples), 20%, 30%, 40%, 50% (5,400 synthetic samples) and only training with synthetic data (100%)

Quantitative Results - Training



Task e: Training the Captioning Model by adding synthetic data to the training set – Quantitative Results (2) and Quantitative Analysis

Quantitative Results - Testing



Quantitative Analysis

- Training with synthetic data resulted in small but consistent improvements, especially in ROUGE-L and METEOR.
- The "best" came from models trained with 20–40% synthetic data, which outperformed the baseline without overfitting to much
- BLEU scores showed no significant gains, and the 100% synthetic model performed worst in terms of generalization, showing clear signs of overfitting
- Overall, synthetic data was beneficial when used in moderation

Task e: Training the Captioning Model by adding synthetic data to the training set - Qualitative Results and Analysis



- GT: Instant Pot Braised Lamb with White Beans and Spinach
- Baseline: - Beans and Beans with Beans Yog and Beans Yog Sauce Yog and Beans Yog Sauce Yogi Yogi
- Plus 10% synthetic: - Beans Lent with and Beans Pe and Beans Yog Sauce Yog Sauce Yog Sauce Yog Sauce Yog and Beans
- Plus 20% synthetic: - Beans Lent with and Beans Pe and Beans Yog Sauce Yog Sauce Yog Sauce Yog Sauce
- Plus 30% synthetic: - Meataf with Beans Goat and Beans Yog Sauce Yog Sauce Yog Sauce Yog Sauce Yog
- Plus 40% synthetic: -ed with and andils beef with andumin andumin paste garlico goatis greens herbs herbs
- Plus 50% synthetic: - Beefender with and Beans Lent and Beans Yog Sauce Yog Sauce Yog Sauce Yog Sauce
- 100% synthetic: ris fried with and sauce lemonah lemonah lemonah vinegar olive oil lemonah oil garlic herbs olive



- GT: Hibiscus Tea Sorbet
- Baseline: - Ice Ice with Ice,,,,,,,,,,,,,
- Plus 10% synthetic: -rozen Ice with Ice and Ice Creamustaze Cream Ice Cream Ice with Ice Cream Ice Cream Ice
- Plus 20% synthetic: -ime ice with and- ice creambet dries ice with ice and ice cream drizzle ice
- Plus 30% synthetic: - Ice Ice with Ice,,,,,,,,,,,,,
- Plus 40% synthetic: - Ice Ice with Ice,,,,,,,,,,,,,
- Plus 50% synthetic: -----
- 100% synthetic: and ice with and creamzzle ice cream strawberryousse ice creamzzle creamzzle creamzzle creamzzle cream

Qualitative Analysis

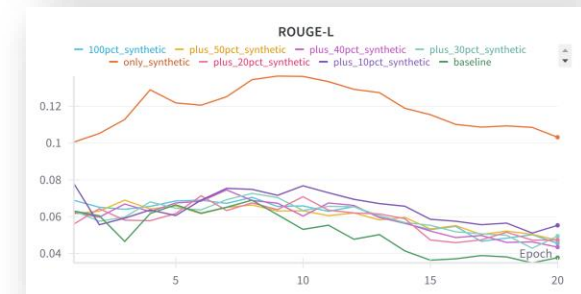
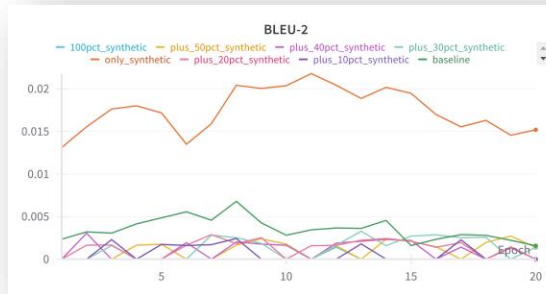
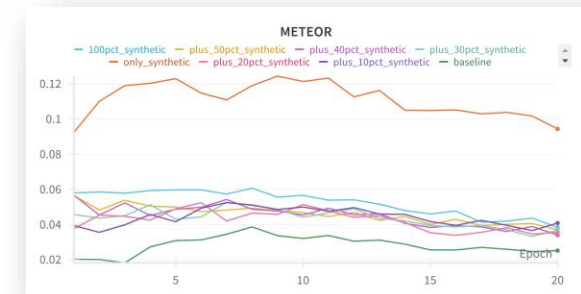
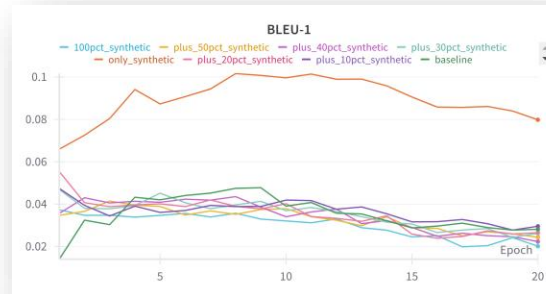
- The baseline model struggles with repetition and incoherent phrases like “Yog Sauce Yog Sauce Yog Sauce”
- As synthetic data is added, predictions become slightly more relevant, with food-related terms appearing more often
- By 30–50% synthetic, the model starts generating ingredients like “beef,” “beans,” and “herbs,” which loosely match the ground truth
- For the second image, the 100% synthetic model outputs terms like “ice cream” and “strawberry mousse,” which, while not accurate, are closer in meaning to “sorbet”
- In short, synthetic data helps reduce repetition and brings predictions closer to the right concept, but full accuracy and fluency remain a challenge

Task e: Training the Captioning Model by only using synthetic data for the training, validation and test set - Idea, Method and Quantitative Results

Idea & Method

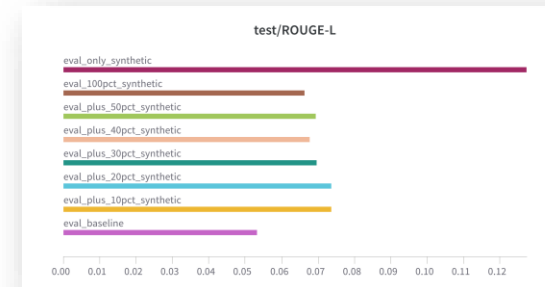
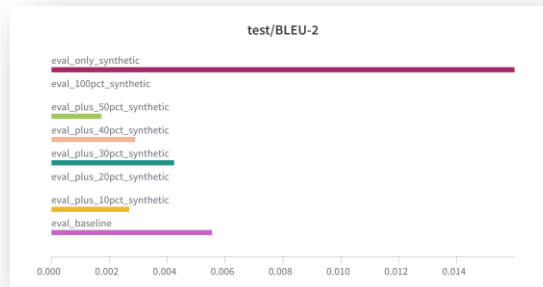
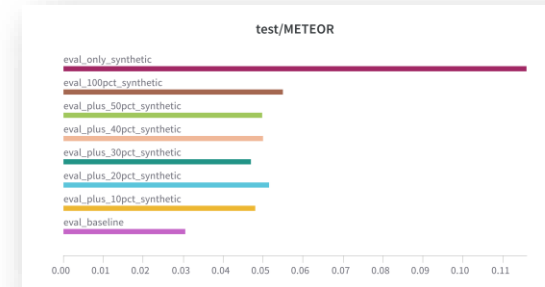
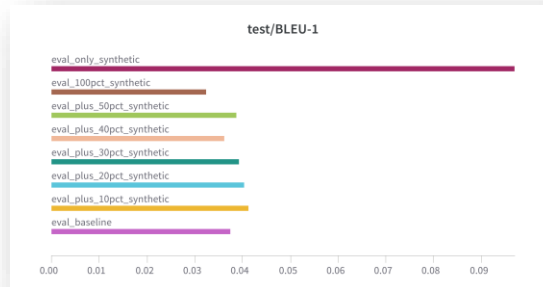
- Since we did not improve our model in comparison to last week, even with 5.400 realistic looking synthetic images, we thought of showing that neither the model nor the synthetic data is the problem but possibly the dataset or more specifically the Titles of the dishes or the ground truth
- Therefore we decided to not only train but also validate and test with only synthetic data
- The reason for that is, that the synthetic data or more specific the generated prompts or Titles are way more descriptive than the original ground truth
- Given this idea we created a whole new data set (split 80/10/10) only consisting of synthetic data consisting of 5.400 generated images and used the same model as well as the same parameters for the training and testing

Quantitative Results - Training



Task e: Training the Captioning Model by only using synthetic data for the training, validation and test set – Quantitative Results (2) and Quantitative Analysis

Quantitative Results - Testing



Quantitative Analysis

- The results show a noticeable improvement across all evaluation metrics compared to the mixed or original datasets
- The model trained and tested on only synthetic data achieved the highest scores in BLEU-1, BLEU-2, ROUGE-L, and METEOR
- These models also show the lowest training and validation loss, suggesting better learning stability and less overfitting
- This supports the idea that the original dataset's labels may be a limiting factor in model performance, and that this synthetic data provides cleaner and more learnable supervision

Task e: Training the Captioning Model by only using synthetic data for the training, validation and test set - Qualitative Analysis



- GT: Muesli with apple slices and almond butter
- Only synthetic: asted and apple with and oats oats oats cinnamon cinnamonins apples cinnamonins oats cinnamonins apples cinnamonaze



- GT: Tofu eggless sandwich with tomato and spinach
- Only synthetic: illed sandwich tomato with and cheesezz cheesezz cheeseioli cheeseolazz breadzz cheeseioli

Qualitative Analysis

- While the predictions from the model trained on only synthetic data are not perfect, they show a clear shift towards structured and food-related phrases.
- In the first example, terms like "oats", "apples", and "cinnamon" appear frequently, showing the model's focus on key ingredients even if the phrasing is repetitive
- In the second case, the prediction includes "sandwich", "tomato", and multiple variations of "cheese", indicating the model's understanding of food components despite some unnatural endings like "cheesezz" or "cheeseolazz"
- These outputs suggest that while the model may over-repeat or invent tokens, the core food concept is preserved more consistently - highlighting that synthetic captions help the model focus on relevant ingredients, though at the cost of good linguistics

Task e: Training the Captioning Model by using synthetic data - Conclusion

Conclusion

- We tried two strategies using synthetic data to improve our captioning model
- One was to add synthetic samples to the original training set in different amounts
- The other was to train and evaluate the model using only synthetic data
- The experiments showed that adding synthetic data can help reduce overfitting and sometimes improve performance. Small and medium amounts seemed to give the most balanced results
- Interestingly, the model trained only on synthetic data (with no original samples) outperformed all other models across all key evaluation metrics
- This was surprising, especially because the total number of synthetic samples was still lower than the full original dataset
- One possible reason is that the synthetic titles are more descriptive and easier to learn from compared to the original dataset
- These results suggest that the quality and clarity of the captions - whether prompts or ground truth – can play a big role in how well the model learns
- This means that improving or rewriting the original titles could also be a promising direction

Team 8 Summary: Synthetic Data Shows Potential, with Caption Quality Likely Playing a Key Role.




Model and Parameter Tuning

Models tested:

Stable-diffusion 2.1
Stable-diffusion 2.1 turbo
Stable-diffusion XL
Stable-diffusion XL turbo

Parameters tested (loop):

DDPM and DDIM
With and without negative prompting
CFG: [7, 12, 17, 22, 27, 32]
Steps: [25, 50, 75, 100, 125, 150, 175]

cfg / steps	20	60
5		
13		

sdxl-turbo
DDPM
with negative prompt
steps=50
cfg=9



Pipeline



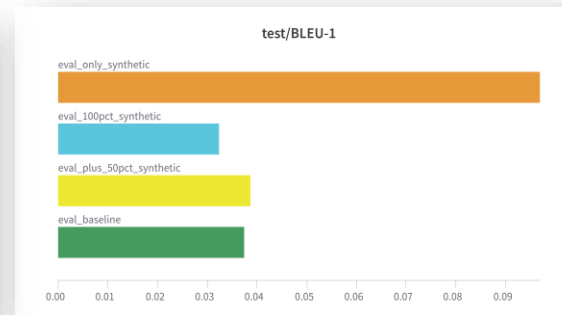
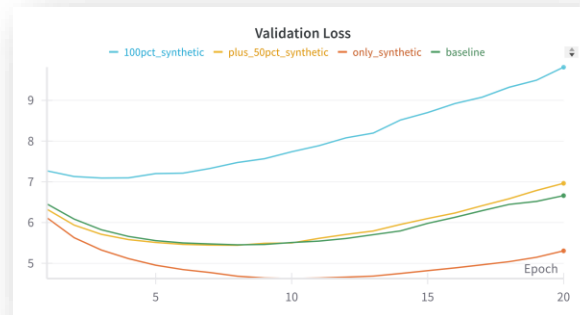
Method

We generated 5.4k synthetic samples to match our 10.8k training set and fine-tuned ViT-GPT2 with 10–50% added synthetic data.

We also trained two models on synthetic data only for comparison, then testing on original and synthetic data



Results



- Like in previous weeks overfitting started around the 10. Epoch and was still pretty strong
- Only minor improvements by adding synthetic data
- Metrics-wise the model trained and tested only on synthetic data performed by far the best – this may be because the captions are more descriptive and the data is cleaner



Predictions



Plus 50% synthetic:
- Beefender with and Beans Lent and Beans Yog Sauce
Yog Sauce Yog Sauce Yog Sauce



Train/test on synthetic:
asted and apple with and oats oats oats cinnamon cinnamonins apples cinnamonins oats cinnamonins apples cinnamonaze