Session 5 – Difussion Models

Team 4:

- María José Millán
- Agustina Ghelfi
- Laila Aborizka

MODELS EXPLORATION

Prompt	sd-turbo	stable-diffusion-2-1	sdxl-turbo	stable-diffusion-xl-base-1.0	stable-diffusion-3.5-large-turbo	stable-diffusion-3.5-medium
“Strawberry Coconut Cake”						
“People walking in the street of a city”						

Initial Model Comparison Using Varied Prompts

In this first analysis, we explored the behavior of various proposed models using different prompts. As seen in the examples, one prompt is food-related ("Strawberry Coconut Cake"), which aligns with a ground truth caption from the dataset, while the other prompt ("People walking in the street of a city") tests the models' performance in a different context.

- Turbo models tend to generate images more quickly but with less detail and lower visual quality compared to their non-turbo counterparts. They create images where the light appears more saturated.
- The XL and 3.5 series models deliver noticeably better performance. For the food-related prompt, both XL and 3.5 models produce visually similar and high-quality outputs. However, when it comes to the street scene prompt, the 3.5 model demonstrates superior detail—especially in rendering people's legs and surrounding architecture—indicating a higher level of visual fidelity.

The subsequent exploration will focus on the non-turbo versions of the models, allowing us to prioritize image quality and detail over generation speed. Additionally, the analysis will center on food-related prompts to align more directly with the dataset we are working with, ensuring a more targeted and relevant exploration.

DDPM VS DDIM

To begin with this exploration phase we generated the images with:

- prompt: "A basket filled with a variety of fresh fruits, soft natural lighting, high detail, photorealistic, 8k resolution"
- num_inference_steps: 50 (default)
- guidance_scale:7.5

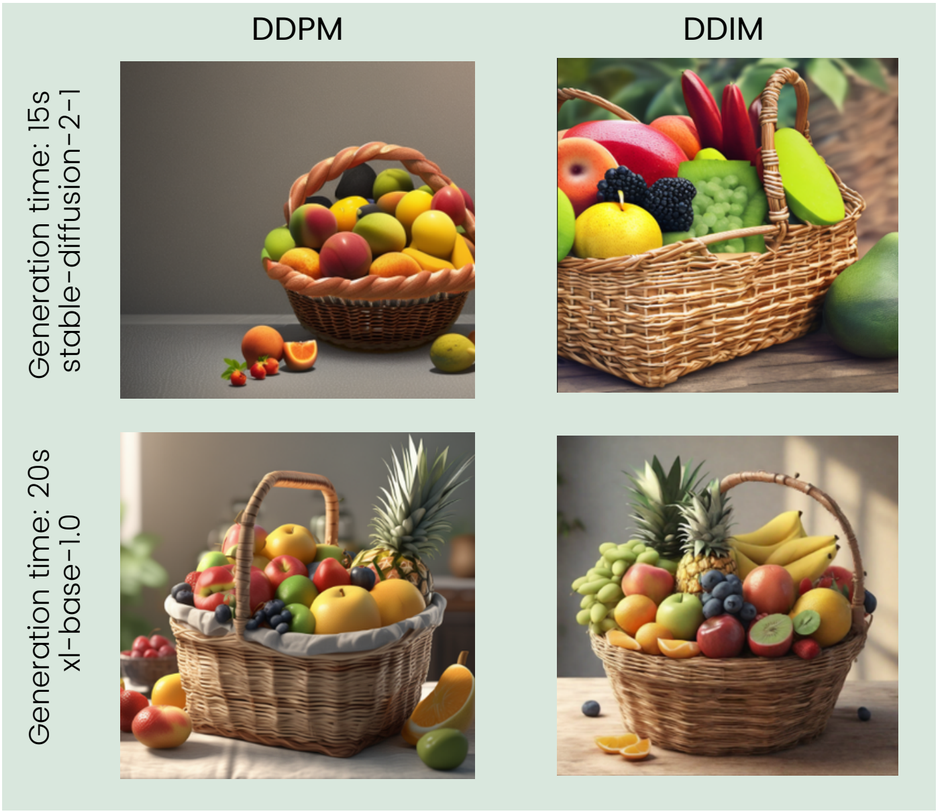
In diffusion models, schedulers control how noise is added and removed during the training and generation process. They define the sequence of noise levels across timesteps, directly impacting the quality and speed of image synthesis.

DDPM (Denoising Diffusion Probabilistic Models):

- A generative model that learns to reconstruct images by progressively removing added noise.
- Requires a large number of steps (up to 1000) to generate high-quality images, leading to longer inference times.

DDIM (Denoising Diffusion Implicit Models):

- An extension of DDPM that introduces a deterministic reverse process, removing the need for random sampling at each step.
- Allows for significantly fewer generation steps (e.g., from 1000 to 50) without sacrificing image quality.
- Enables faster and more controllable image generation, while maintaining high fidelity.



DDIM appears to outperform DDPM in both models, as it produces high-quality images. The generated examples show sharper details and better overall fidelity, highlighting DDIM's efficiency and speed advantage without sacrificing visual quality.

Stable Diffusion 3.5 adopts a newer scheduler: FlowMatchEulerDiscreteScheduler. This scheduler, inspired by flow-based generative models, enables even sharper image details, improved coherence, and faster generation in certain scenarios.

Generation time: 88.70s
3.5-medium

POSITIVE & NEGATIVE PROMPTING

A **positive prompt** is a standard text prompt provided to the model. It describes the subject, style, composition, and other desired elements of the image you want to generate.

A **negative prompt** is additional text prompt that specifies elements, styles, or qualities you explicitly want to avoid in the generated image.

The only change in relation to the last configuration is the "negative_prompt": "apples, pears, apple, pear, malformed fruits, blurry, low quality, cartoon, CGI, oversaturated, unrealistic proportions."

We consider that the best results were given by xl model so the next parameters exploration will be with this model.

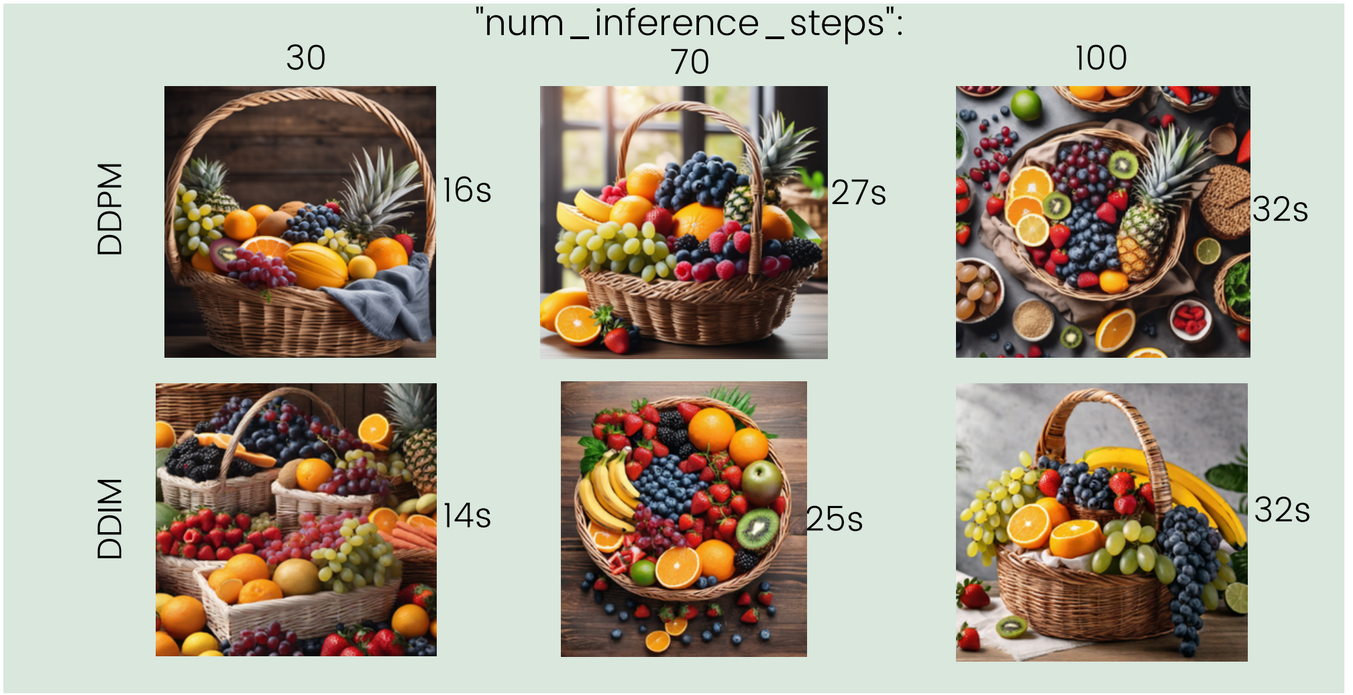

3.5-medium

NUMBER OF DENOISING STEPS

The **num_inference_steps** parameter defines how many times the denoising process is repeated during image generation. Each step progressively refines the image by removing noise based on the model's training and the given prompt (especially when CFG is enabled). Stable Diffusion performs well with around 50 steps by default, offering a good balance between speed and quality.

- More steps generally lead to higher-quality images, but also increase generation time.
- Reducing the number speeds up generation, while increasing it may improve visual fidelity.

As observed, increasing the number of denoising steps leads to noticeably better image quality and detail, especially in the textures and shapes of the fruits. However, this improvement comes at the cost of longer generation times.

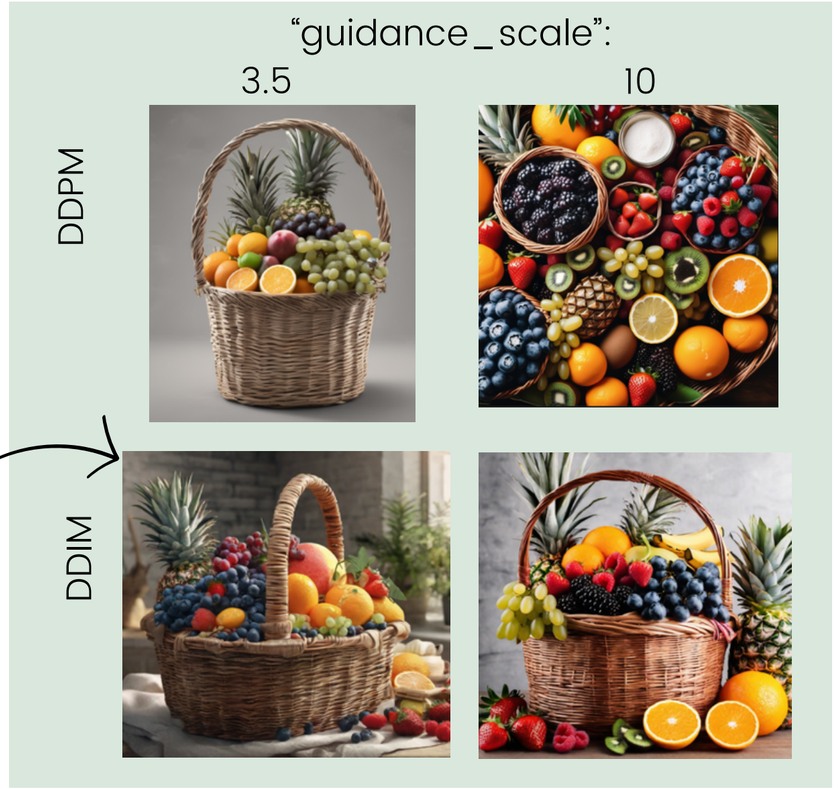


STRENGTH OF CFG

Classifier-Free Guidance (CFG) is a technique used in diffusion models to improve how closely generated images match the input prompt.
This is controlled by the guidance_scale parameter, which adjusts the strength of the guidance.

During inference, the model combines predictions from both a conditioned (with prompt) and an unconditioned (without prompt) version of the model. This helps steer the generation toward the desired output.
Higher values of guidance_scale increase prompt adherence, but may reduce image diversity or introduce artifacts if set too high.

There is an apple in the basket, not following the prompt stricly



CONCLUSIONS

Throughout this parameter exploration stage, we analyzed key components that influence the image generation process in diffusion models, including the choice of scheduler (DDPM vs. DDIM), the number of denoising steps, the use of classifier-free guidance (CFG), and the role of both positive and negative prompts. Our experiments showed that DDIM consistently outperformed DDPM in both quality and speed, making it a more efficient and reliable choice. Increasing the number of denoising steps improved image detail—especially in complex textures like fruits—while CFG helped strengthen alignment with the input prompt, though excessive values reduced diversity. Negative prompting also proved useful in filtering out unwanted content and refining visual output.
Based on the balance between image quality, prompt adherence, and generation time, the final configuration chosen to continue with the analysis was:

- Model: stabilityai/stable-diffusion-xl-base-1.0
- Prompt: "A basket filled with a variety of fresh fruits, soft natural lighting, high detail, photorealistic, 8k resolution"
- Negative Prompt: "apples, pears, apple, pear, malformed fruits, blurry, low quality, cartoon, CGI, oversaturated, unrealistic proportions"
- Scheduler: DDIM
- Number of inference steps: 80
- Guidance Scale: 8

PROBLEM FORMULATION

1 Dataset Exploration

Using **spaCy's 'en_core_web_lg'** pipeline, we extracted and analyzed food terms from image captions

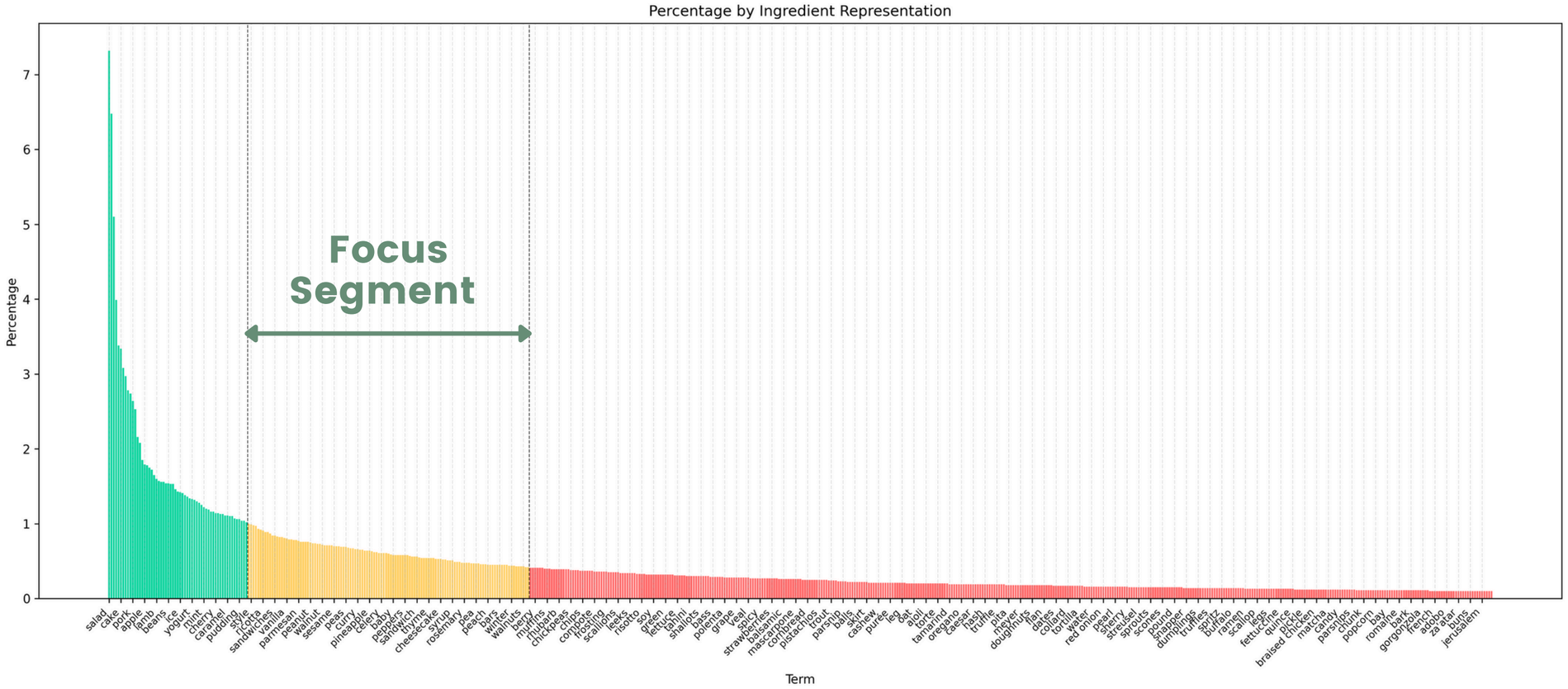
2 Term Classification

We classify our food terms into 3 classes:

- **Common (>1% frequency)**: Dominant ingredients
- **Underrepresented (0.4-1%)**: Less frequent ingredients (Focus Segment)
- **Discarded (<0.4%)**: Rare terms

3 Problem Formulation

How does **targeted oversampling of underrepresented ingredients using Diffusion Models** affect model predictions in an image captioning task?



- Common Appear in over 130 images
- Under-represented ★ Targeted Oversampling
- Discarded We discarded ingredients with less than 52 images

GENERATION PIPELINE

The goal of this pipeline is to automatically generate high-quality, photorealistic images along with captions of food dishes containing specified ingredients, specifically for training our food captioning model.

Ingredient Selection

- Number of ingredients selected: 87
- Selection criteria: Ingredients that belong to our predefined focus segment
- Image generation per ingredient: 20 images each
- Total images generated: 1,740 photorealistic images

Prompt Engineering (LLM Phase)

- Utilizes meta-llama/Llama-3.2-1B-Instruct, a compact LLM for generating short, vivid food image prompts.
- Ensures each prompt includes a specific ingredient in a dish context rather than as a standalone item.
- Prompts include details about cooking style, plating, and food presentation.

Image Generation (Diffusion Model)

- Model: stabilityai/stable-diffusion-xl-base-1.0
- Prompt: from LLM
- Negative Prompt: "deformed, unappetizing, incorrect texture, malformed , blurry, low quality, cartoon, CGI, ,..."
- Scheduler: DDIM
- Number of inference steps: 80
- Guidance Scale: 8

Caption Generation

- Qwen2.5-VL to generate captions for the 1,740 synthetically generated images.
- It handles complex visual inputs and generates coherent titles, improving the realism and usefulness of our synthetic dataset.
- It address the common issue of non-descriptive or generic image captions

RESULTS

Examples of synthetic food images and captions generated by our pipeline. Each image was generated from a prompt specifying a key ingredient, and subsequently captioned using Qwen2.5-VL. These samples were selected to illustrate both the strengths and limitations of the generation process.


Grilled Zucchini with Garlic and Herbs


Zucchini Fritters with Lemon and Basil


Zucchini Lasagna with Basil and Cheese


Teriyaki Chicken and Zucchini Stir-Fry Bowl


Asparagus and Zucchini Boats with Creamy Sauce


Zucchini Salad with Almonds and Basil

ZUCCHINI-BASED PROMPTS

The model successfully produced a diverse array of realistic food dishes where zucchini is not only present, but meaningfully integrated into the recipe (e.g., fritters, salad).

These results demonstrate strong ingredient-context integration and visual realism.


Indian Butter Chicken with Naan and Coriander


Chicken Tikka Masala with Cilantro Rice


Grilled Pineapple Skewers with Curry Dip


Chicken Tikka Masala with Basmati Rice

CURRY-BASED PROMPTS

In contrast, the curry-based outputs lack diversity—nearly all generated images depict chicken curry in various presentations (with rice or naan).

Despite different prompts, the model fixated on a single dish archetype, reducing the augmentation's variability.

Low Caption Variance

Even though the images are photorealistic, the captions are often too similar or predictable, especially when the generated dishes are repetitive (as in the curry case).

Bias Reinforcement

The generation model might reinforce existing biases in food image-caption pairs (e.g., "curry = chicken curry"), thus failing to inject new concepts into the dataset.

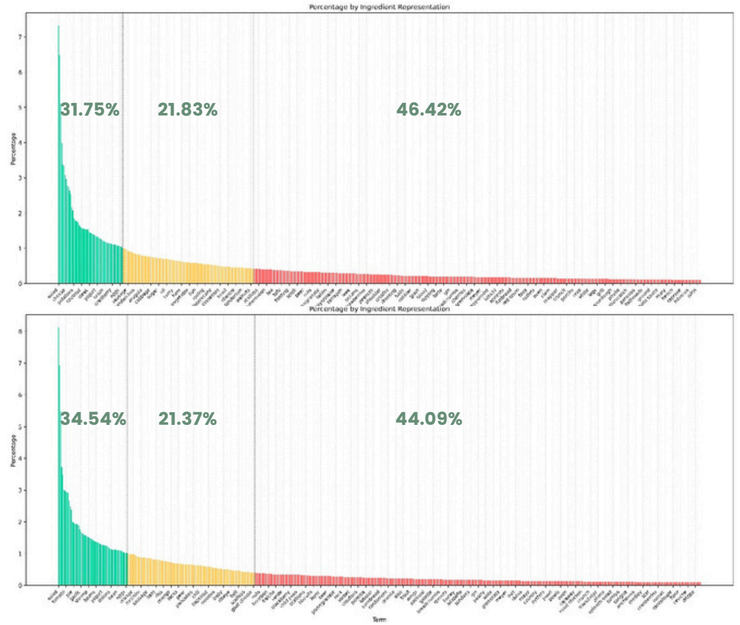
Synthetic Realism Gap

It's not just about making images that look realistic—what really matters is whether they introduce new and varied patterns the model can learn from. If the synthetic data is too repetitive or too similar to what's already in the training set, the model may not gain much from it, even if the images appear high-quality.

RESULTS

IMPACT OF AUGMENTATION OVER THE DATASET

The plots below show the distribution of ingredient frequency before (top) and after (bottom) the augmentation process. The green bars represent “common” ingredients, while the yellow bars correspond to “underrepresented” ones. Our main objective with this augmentation was to reduce the long-tail imbalance and achieve better representation of underrepresented ingredients.



After augmentation, several ingredients such as avocado moved from underrepresented to common. However, some ingredients, such as pudding, shifted in the opposite direction—becoming underrepresented after augmentation.

This illustrates that while the augmentation helped smooth the distribution in some areas, it also unintentionally reinforced imbalance in others. This may be a consequence of our filtering step, which excluded many extremely rare ingredients (those with only one appearance), even though their cumulative presence forms a significant portion of the dataset.

PERFORMANCE EVALUATION

The table on the right compares model performance between last week's Llama 3.2 3B model and this week's version after applying the image+caption augmentation.

Although we observe improvement across all metrics, the gains are subtle:
BLEU-1: +0.008
METEOR: +0.012
ROUGE-L: +0.02

These improvements suggest a minor benefit from the augmentation strategy, but not enough to conclude that it had a significant positive impact on the model’s captioning ability.

	last week	this week
BLEU_1	0.269	0.277
BLEU_2	0.141	0.150
ROUGE_L	0.228	0.248
METEOR	0.160	0.172

CONCLUSIONS

While the synthetic samples were visually promising and led to a modest improvement in ingredient balance, the augmentation did not result in a meaningful boost in captioning performance.

Despite the realism of the generated samples, limitations such as repetitive dish types, lack of linguistic diversity, and minimal novelty in content likely prevented the augmented data from providing meaningful improvements.

We also explored cuisine-level imbalance, finding many underrepresented styles (e.g., French, Mexican, Chinese). While our initial intuition was that cuisine might be less informative than ingredients, we now believe combining both could yield more balanced and diverse augmentation.

Moreover, many ingredients were excluded from augmentation due to low occurrence. Even though each of these is rare, together they represent a substantial portion of the long tail. Future work could focus on clustering rare ingredients into conceptual groups for more scalable augmentation.

SUMMARY

EXPLORATION

Models: sd-turbo, stable-diffudision-2-1, sdxl-turbo, stable-diffusion-xl-base-1.0, stable-diffusion-3.5-large-turbo, stable-diffusion-3.5-medium

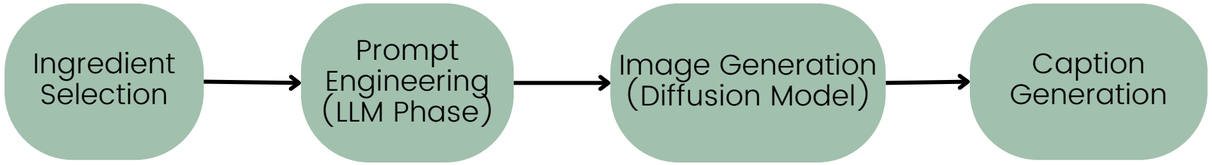
Findings:

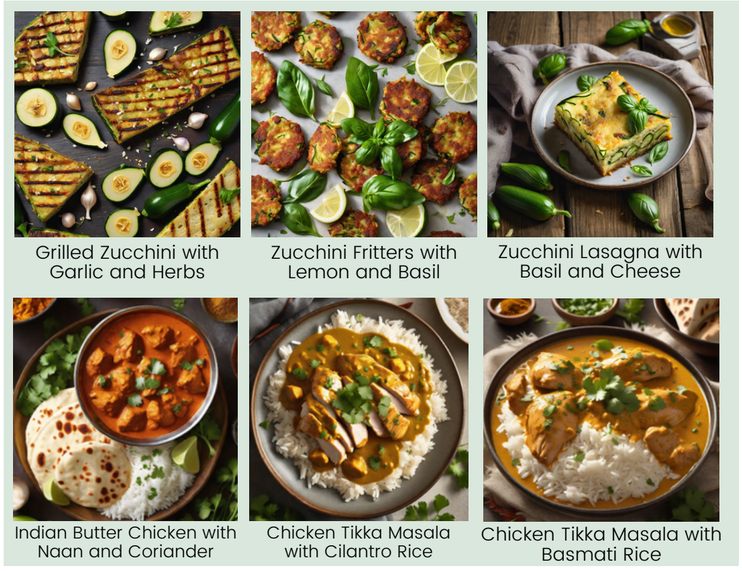
- Sheduler: DDIM is more effective than DDPM.
- More denoising steps improved detail.
- Guidance techniques enhanced prompt alignment and output quality.
- negative prompts helped suppress unwanted elements.

Chosen configuration based on the balance between image quality, prompt adherence, and generation time:

- Model: stabilityai/stable-diffusion-xl-base-1.0
- Include negative prompts.
- Scheduler: DDIM
- Number of inference steps: 80
- Guidance Scale: 8

IDENTIFY A PROBLEM AND GENERATE SYNTETHIC SAMPLES

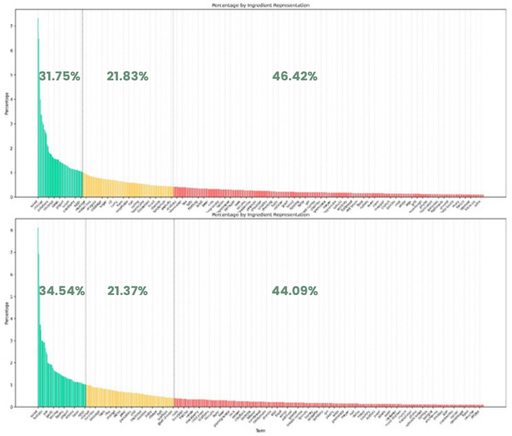




Some varied results

Some results were repetitive

GOAL:
Reduce the long-tail imbalance and achieve better representation of underrepresented ingredients (yellow).



PERFORMANCE OVER CAPTIONING MODEL

These improvements suggest a minor benefit from the augmentation strategy, but not enough to conclude that it had a significant positive impact on the model's captioning ability.

	week 4 3.2 3B	week 5 3.2 3B
BLEU_1	0.269	0.277
BLEU_2	0.141	0.150
ROUGE_L	0.228	0.248
METEOR	0.160	0.172