

Multimodal Recognition Project

Final Session: Project Presentation

Group 6

Module: C5 - Visual Recognition

Coordinators: Ernest Valveny, Carlos Boned, Lei Kang

Abril Piñol

Biel González

Mireia Majó

Alejandro Donaire

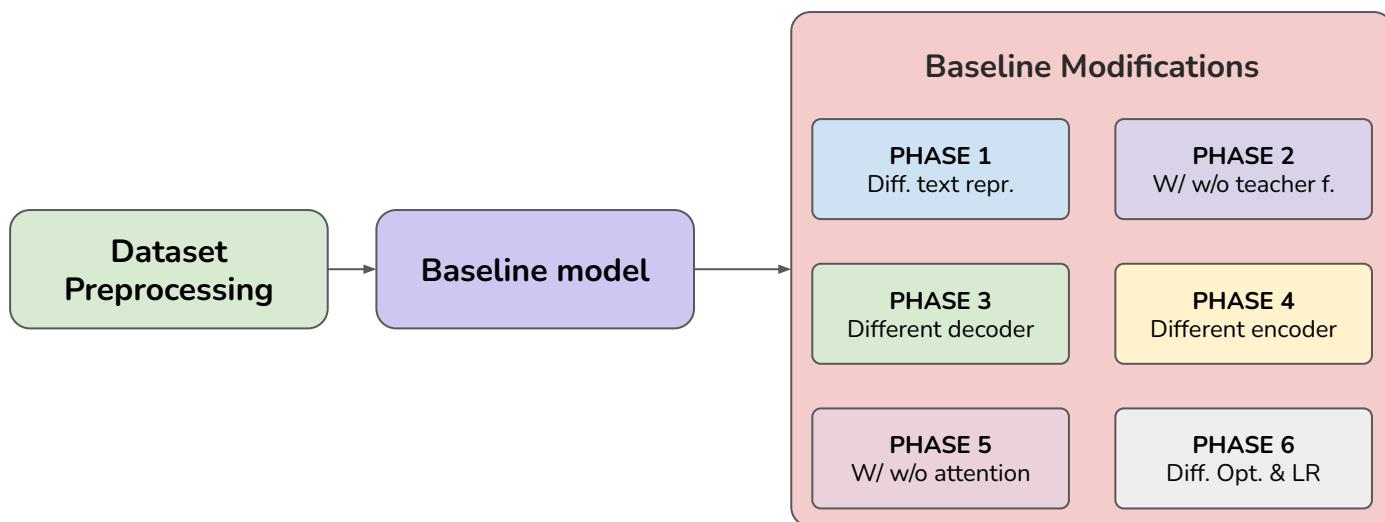
Contents

1. **Session 3: Image Captioning (1)**
 - 1.1. Data analysis, processing, and split
 - 1.2. Baseline model
 - 1.3. Baseline modifications
 - 1.4. Experiments overview
 - 1.5. Results
2. **Session 4: Image Captioning (2)**
 - 2.1. ViT-GPT2
 - 2.2. LoRA-tuned LLMs
 - 2.3. SoTA LLMs
 - 2.4. Model Comparison
3. **Session 5: Diffusion models**
 - 3.1 Exploring SD
 - 3.2 Identifying problems
 - 3.3 Generating Synthetic Data
 - 3.4 Retraining the model
 - 3.5 Why did we get mixed results?
4. **Conclusions and Insights**

1

Session 3: Image Captioning (1)

Workflow overview for this session:



Data analysis



NaNs



Duplicates

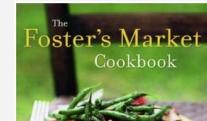


Non-overlap



Rare char.

Min: 3
Max: 112



Strange images

Different lengths

Preprocessing



13,466
Valid
(img, cap.)

Data split

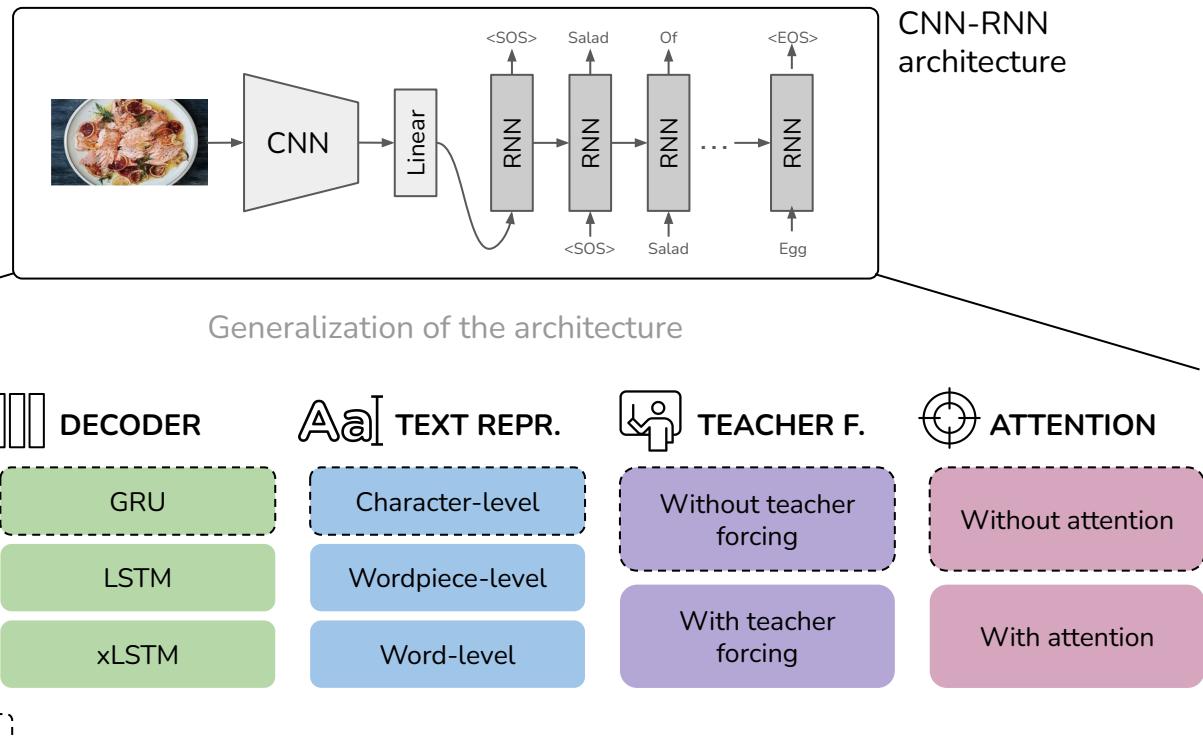


Train
80

Val.
10

Test
10

Baseline model modifications



Baseline model modifications (details)



ENCODER

- **Projection** before passing features to decoder (always 512 dim.).
- **Pooling** before extracting final VGG features to reduce dim.



DECODER

- In **xLSTM**: introduction of third state, extended memory gate, and layer normalization to the hidden, cell, and memory states.



TEXT
REPRESENTATION

- **Characters**: list of common characters.
- **Words**: top N most common words.
- **Wordpiece**: autotokenizer.



TEACHER FORCING

- Modifications to **forward function & decoding loop** to add GT tokens.
- Can choose **teacher forcing ratio** (0.8) and **teacher forcing decay** (0.95).

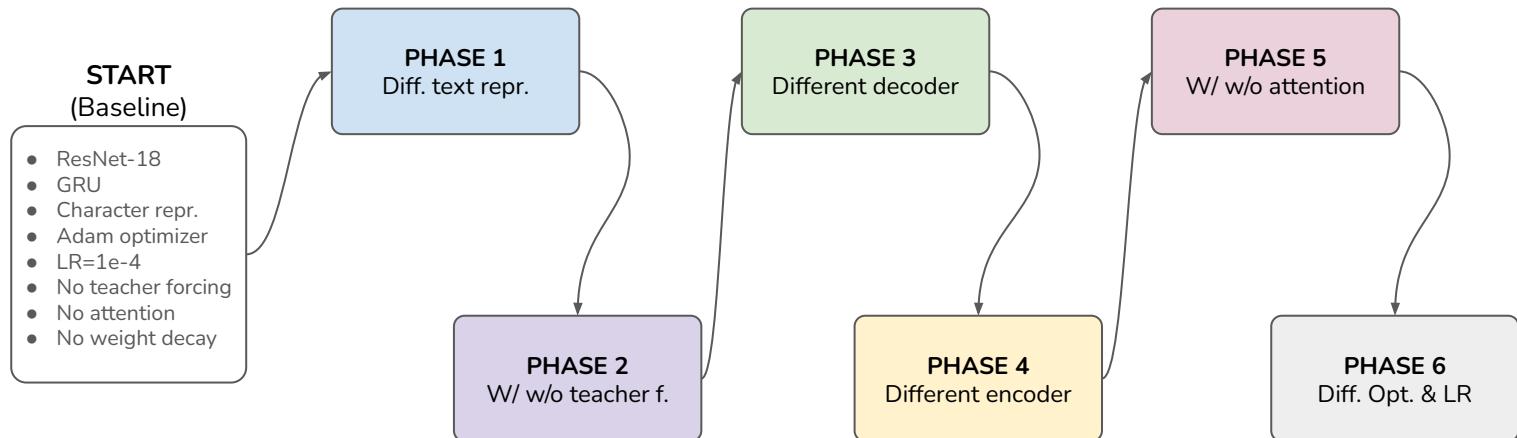


ATTENTION

- **Bahdanau attention** (additive attention).
- Keep **feature maps unflattened** so decoder can attend over them.

Experiments overview

Progressively accumulate good-performing configurations executing experiments in decreasing order of expected importance.



Results

↑ **Text representation:** by far the most impactful in the results.

Text repr.	BLEU-1	BLEU-2	ROUGE-L	METEOR
Character	0	0	0	0
Wordpiece	8.23	0	8.97	4.30
Word	10.35	0	12.78	6.03

↑ **Teacher forcing:** marginal improvements.

≡ **Decoders:** minimal differences between them.

≡ **Encoders:** ResNets outperform VGGs. Depth not important.

✗ **Attention:** slight decrease in performance.

≡ **Other hyperparam.:** no improvements after optimization.

Best: ResNet-18 · LSTM · Word-Lvl · Teacher F. · No Att.

Model	BLEU-1	BLEU-2	ROUGE-L	METEOR
Best	10.37	0	13.07	6.26

Character



Original caption: Big-Batch Pancake and Waffle Mix

Predicted caption: Caaeee

Wordpiece



Original caption: Grilled Radicchio Salad with Sherry-Mustard Dressing

Predicted caption: qrilled with with with with and and and

Word



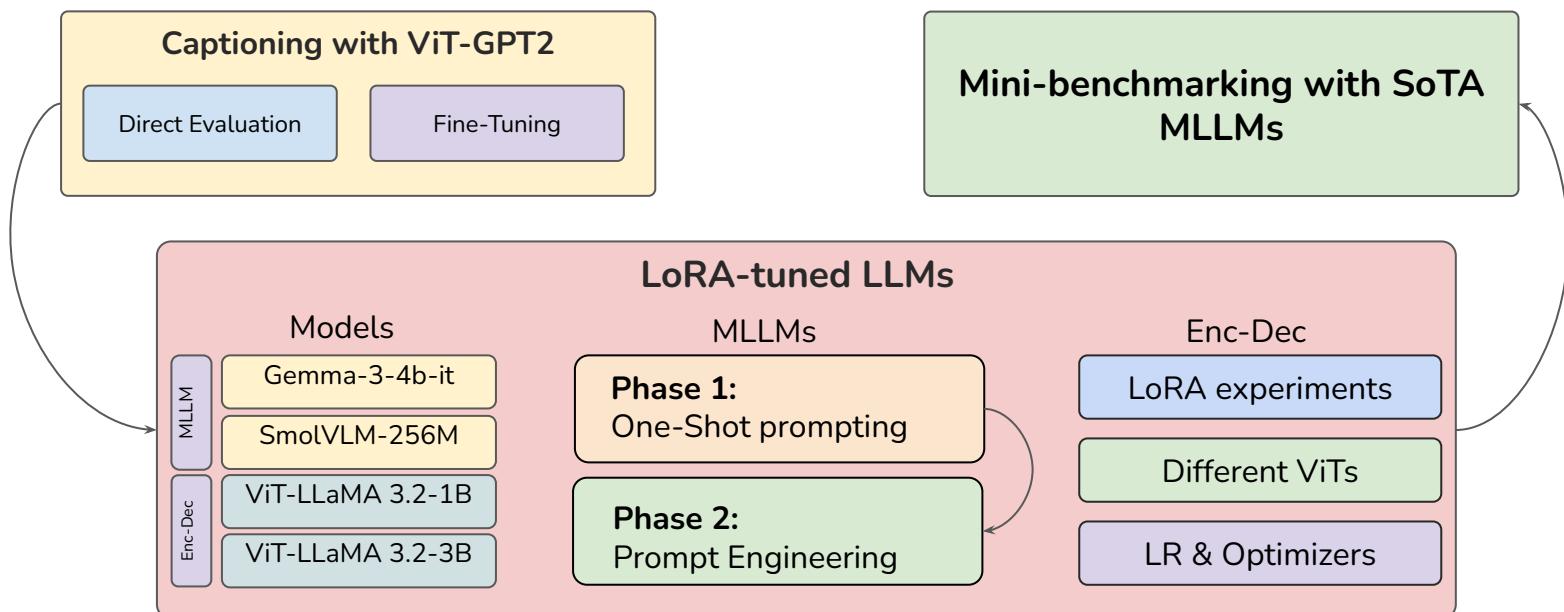
Original caption: Carrot-Coconut Soup

Predicted caption: creamy soup soup soup

2

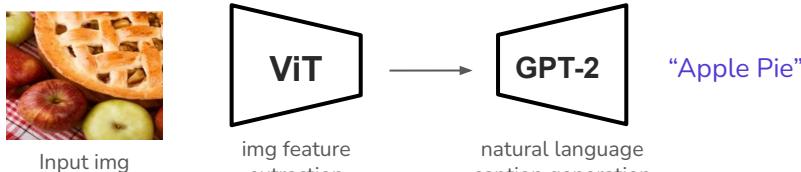
Session 4: Image Captioning (2)

Workflow overview for this session:



Captioning with ViT-GPT2

Architecture



Fine-Tuning

Set-up

Epochs: 25
 Batch size: 16
 Learning rate: 5e-5
 Loss: Cross-entropy
 Evaluation metrics: BLEU-1, BLEU-2, ROUGE-L, METEOR
 Optimiser: AdamW with linear learning rate scheduler

Strategies

ViT (🔥) & GPT-2 (🔥)

ViT (🔥) & GPT-2 (📦)

ViT (📦) & GPT-2 (🔥)

Approach (best config.)	BLEU-1	BLEU-2	ROUGE-L	METEOR
Pre-trained Weights	3.44	0.23	5.80	4.06
ViT (🔥) GPT-2 (📦)	7.37	1.27	8.58	10.74
ViT (📦) GPT-2 (🔥)	14.80	6.22	8.96	13.22
ViT (🔥) GPT-2 (🔥)	16.21	6.54	9.98	14.44

Qualitative results of ViT-GPT2

Image	Original caption	Pre-trained Weights	ViT (🔥) GPT-2 (🧊)	ViT (🧊) GPT-2 (🔥)	ViT (🔥) GPT-2 (🔥)
	Chocolate Panna Cotta Layer Cake	A piece of chocolate cake sitting on top of a plate	A chocolate cake topped with whipped cream cheese frosting	Chocolate Cake with Chocolate Frosting	Chocolate-Hazelnut Cheesecake

More specific captions

More relevant words

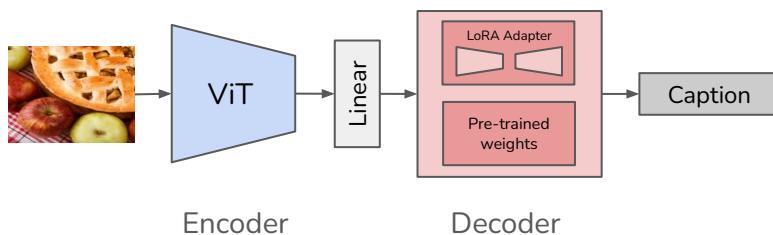
Preparations and dishes

Best results in full
Fine-tuning

Easy to implement and
Fine-tune

Fine-tuning with LoRA

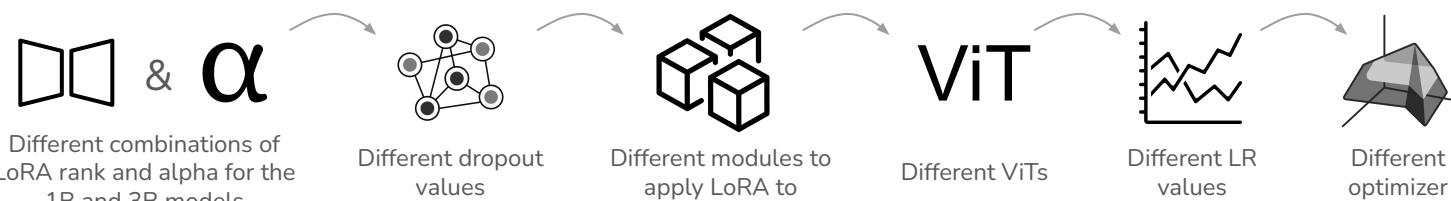
ARCHITECTURE OVERVIEW



IMPLEMENTATION DETAILS

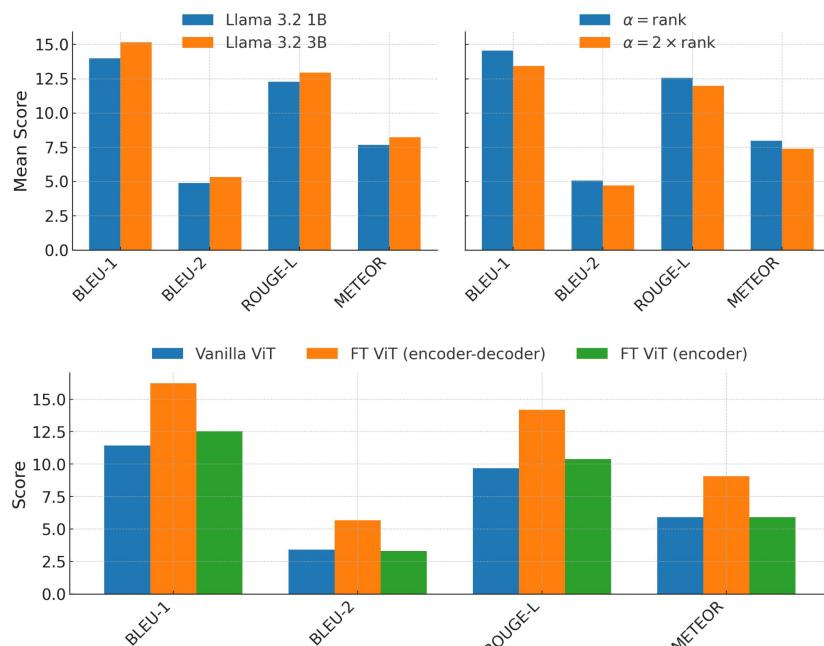
- Trainable **projection** to map ViT and LLaMA hidden sizes.
- **Concatenate** image and text embeddings in the forward pass.
- While training, **teacher-forced** validation.

EXPERIMENTS OVERVIEW



Fine-tuning with LoRA

KEY RESULTS



Insights from the plots:

- LLaMA 3B generally outperforms LLaMA 1B.
- Setting **alpha=rank** works better than following the rule of thumb $\alpha=2 \times \text{rank}$
- Fine-tuning the ViT increases performance.
- FT ViT using enc-dec outperforms the one FT only enc in ViT-GPT2 architecture.

Other insights:

- Ranks 8/16 and alpha 8/16 seem to provide consistently high results.
- The rest of the parameters marginally change the final performance.

Best configuration:

Approach	BLEU-1	BLEU-2	ROUGE-L	METEOR
ViT-LLaMA (3B)	15.69	5.13	13.21	8.25

Mini benchmarking SoTA MLLMs

- Small additional study to grasp how good available SoTA MLLMs are.
- Evaluation of two popular MLLMs:  ChatGPT 4o (April) &  Claude 3.7 Sonnet (April)
- **Prompt:** “Provide the precise culinary name of this dish in under 10 words. Do not include any other information.”
- Manually introducing **10 randomly selected images**.

RESULTS SUMMARY

Metrics	ChatGPT 4o	Claude 3.7 Sonnet
BLEU-1	16.92	12.07
BLEU-2	7.84	0
ROUGE-L	31.35	21.66
METEOR	27.42	18.24



Original: Chinese Egg Noodles with Smoked Duck and Snow Peas
ChatGPT: Asian-style stir-fried noodles with vegetables and ham
Claude: Lo Mein with vegetables

- Competitive results with our fine-tuned models.
- Grammatically perfect sentences but fail to capture fine-grained details.
- Sometimes captions not in the same “style”.

Model comparison

Approach	BLEU-1	BLEU-2	ROUGE-L	METEOR
Worst CNN-RNN (baseline)	0	0	0	0
Best CNN-RNN	10.37	0	13.07	6.26
Best ViT-GPT2 (Enc-Dec FT)	16.21	6.54	9.98	14.44
Best MLLM (Gemma-3-4b)	25.15	12.17	24.19	18.12
Best ViT-LLaMA (3B)	15.69	5.13	13.21	8.25
ChatGPT 4o (n=10)	16.92	7.84	31.35	27.42
Claude Sonnet 3.7 (n=10)	12.07	0	21.66	18.24

- LLM-based approaches far superior to classic CNN-RNN ones.
- Directly prompting MLLMs provided better results than fine-tuning.
- Advanced online SoTA MLLMs like ChatGPT and Claude can compete with our models.

Worst RNN-CNN-based
(ResNet18+GRU)



Original: Big-Batch Pancake and Waffle Mix
Pred: Caaeee

Best RNN-CNN-based
(ResNet18 + LSTM)



Original: Carrot-Coconut Soup
Pred: creamy soup soup soup

Best Advanced MLLM
(ChatGPT 4o)



Original: Chinese Egg Noodles with Smoked Duck and Snow Peas
Pred: Asian-style stir-fried noodles with vegetables and ham

Worst LLM-based
(Pre-trained ViT-GPT2)



Original: Rice with Parsley, Almonds, and Apricots
Pred: A plate of food on a table

Intermediate LLM-based
(LLaMA 3.2 1B)



Original: Chocolate Panna Cotta Layer Cake
Pred: Chocolate Cake with Chocolate Ganache and [...]

Best LLM-based
(Gemma-3-4b)



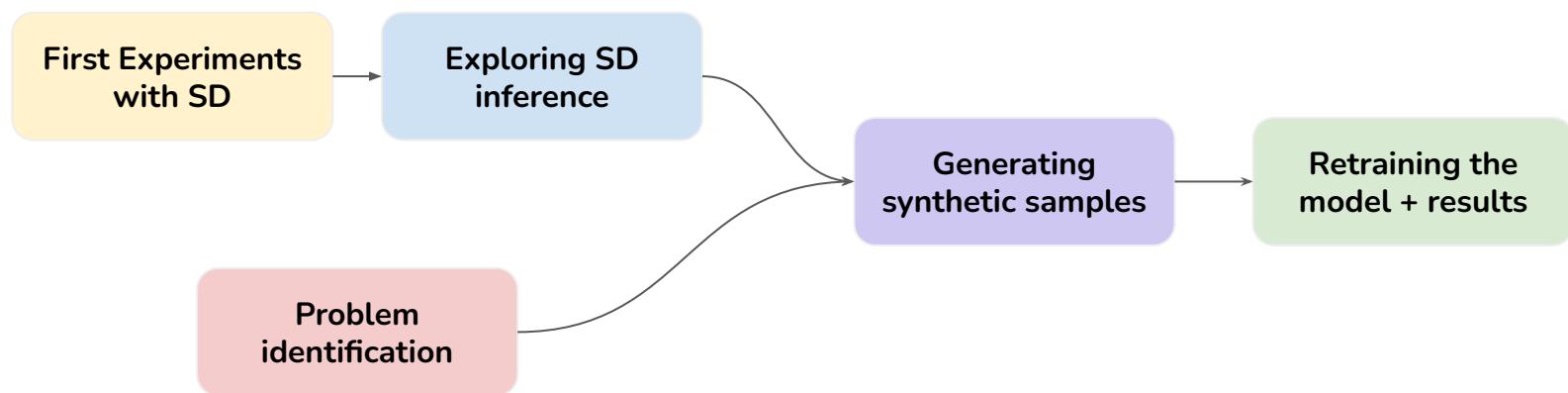
Original: BBQ Pork Chops with Herb-Butter Corn and Sweet Potatoes
Pred: Smoked Ribs with Sweet Potatoes and Corn on the Cob

- CNN-RNNs show rudimentary understanding of image content but lack caption structure, mostly repeating common tokens.
- LLM-based show better caption structure and understanding but struggle with details.

3

Session 5: Diffusion Models

Workflow overview for this session:



Exploring SD - Comparison Setup

For each of these models:

SD 2.1

SDXL Base

SDXL Turbo

SDXL w/ Refiner

SD 3.5 Medium

We generate:

- 4 img w/ rand. seed
- 1 img w/ fixed seed
→ seed 42



We do it for:

- 20 different prompts from FIRD captions



"Apple pie"



"Aperol spritz"



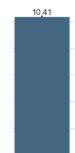
"Halibut confit with leeks, coriander and lemon"

Exploring SD - Analysis

SD 2.1

Gen. time:
1.85s

SDXL Base

Gen. time:
10.41s

SDXL Turbo

Gen. time:
4.41s

SD 3.5 Medium

Gen. time:
14.22s

SDXL w/ Refiner

No significant difference with SDXL Base & high gen. time of 14.00s

Exploring SD - Final models chosen

Parameter optimization
based on **quality** not
metrics

Classifier Free Guidance Scale



Denoising steps



Neg. prompt



Extra experiments

Playing w/ prompt



Apple Pie

Apple Pie, food photography

prompt +
“, food
photography”

Generation size



Best models

Best config.

SDXL BASE

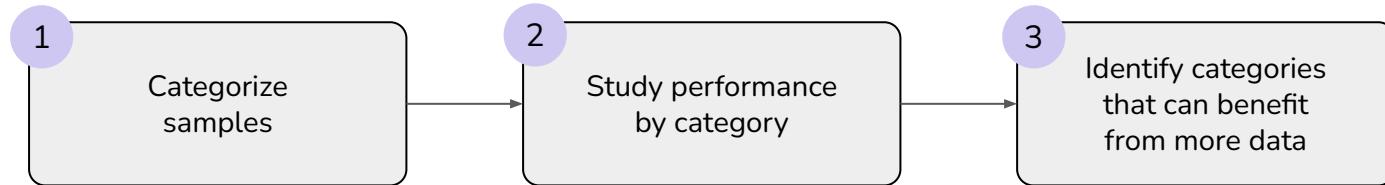
CFG
5Denoising Steps
50Sampler
DDIMNegative Prompt
None

SD 3.5 Medium

CFG
5Denoising Steps
50Sampler
DDIMNegative Prompt
None

Problem identification

GENERAL STRATEGY



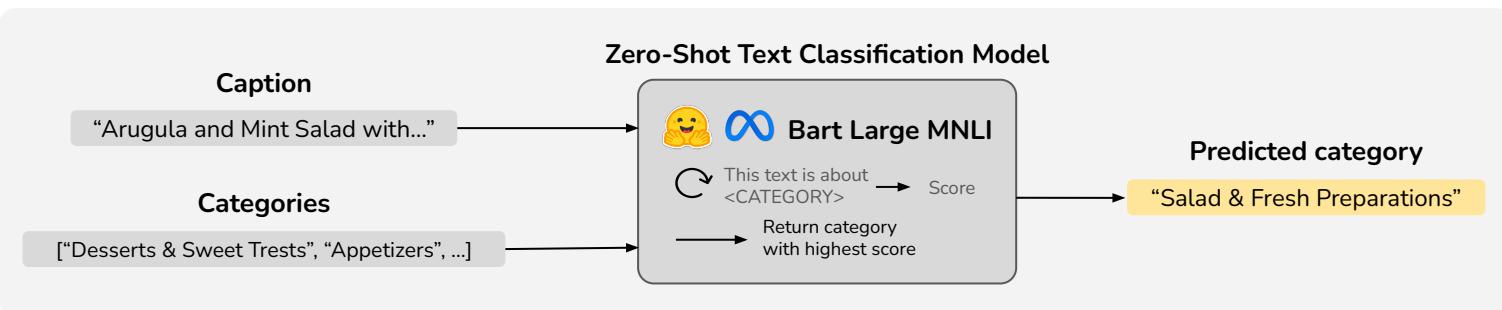
- Helps us **better target our limited time and compute resources.**
- **Easier to evaluate** whether our efforts succeeded by focusing on specific categories.

Problem identification

CREATING THE CATEGORIES

- Feeding **1000 random samples to Claude**. Ask to suggest categories.
 - Posterior **manual refinement** of the suggestions
- } **11** different dish categories

CATEGORIZING THE SAMPLES (THROUGH THEIR CAPTIONS)



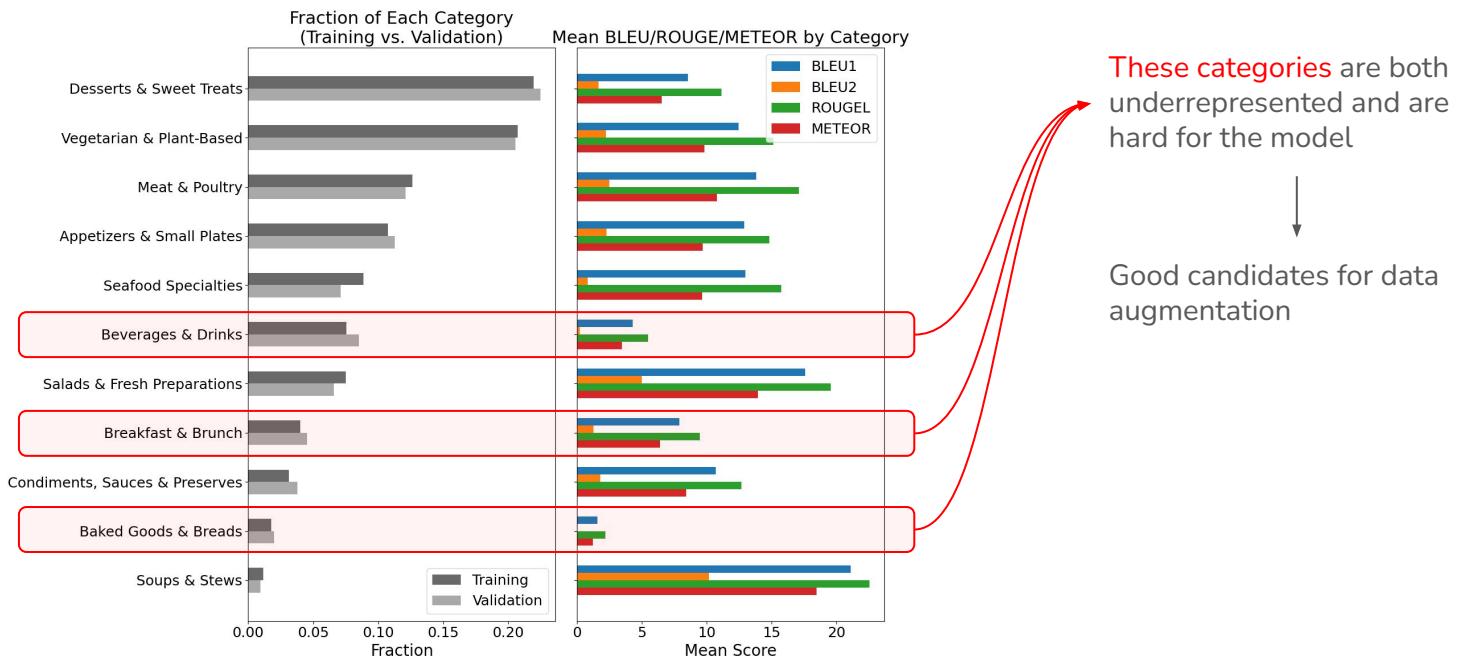
CREATING NEW CAPTIONS

- Feeding **1000 random samples to Claude**. Ask to summarize **caption style** in a prompt.
- Repeatedly **prompt ChatGPT** for new captions. **Check uniqueness** with Python script.

Problem identification IDENTIFYING CATEGORIES THAT CAN BENEFIT FROM MORE DATA

Model that we use:

ViT-LLaMA 3.2 1B with opt. params



Generating synthetic samples to mitigate our problem

For these 3 selected categories:

Beverages & Drinks

Breakfast & Brunch

Baked goods & Breads

We generate **955** different captions

We generate **3** images for each caption



getting **2864** (image, caption) pairs

For each model

SDXL BASE



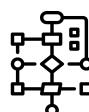
SD 3.5 Medium



Prompt: "Almond-Stuffed Brioche Twists"

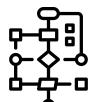
Generating synthetic samples to mitigate our problem

DIFFUSION MODELS USED



Stable Diffusion 3.5 Medium

Optimized for our case of food generation



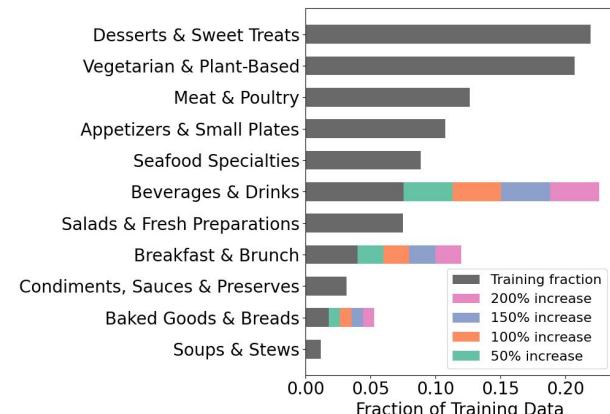
Stable Diffusion XL Base 1.0

Less good than 3.5 Medium, for comparison

EXPERIMENTS OVERVIEW (FOR EACH CATEGORY)

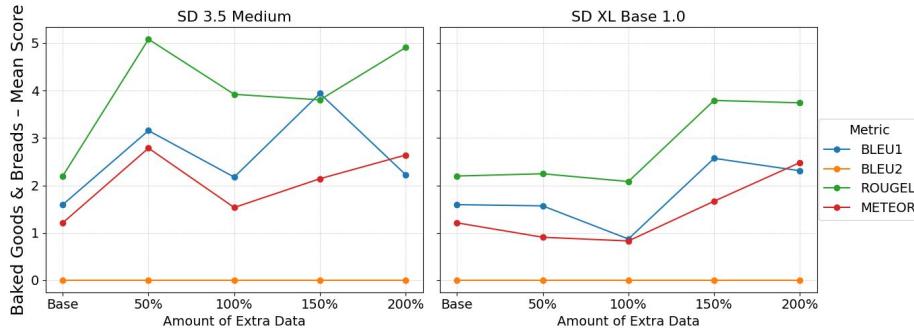


AMOUNT OF DATA GENERATED



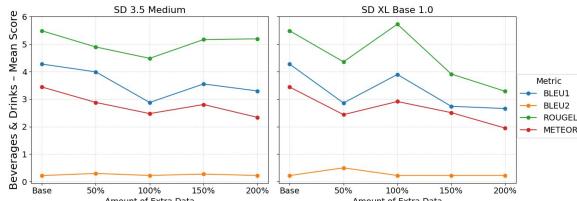
Retraining our model with augmented data

RESULTS FOR “BAKED GOODS & BREADS”

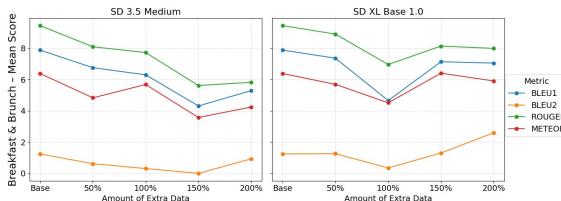


- Only improvements in category “Baked Goods & Breads”.
- More data seems to be better.
- SD 3.5 better than SD XL.
- Other categories are negatively affected.

RESULTS FOR “BEVERAGES & DRINKS”



RESULTS FOR “BREAKFAST & BRUNCH”



Why did we get mixed results?



Model architecture

The model itself might have architectural limitations. Other designs might be more effective.



Data quality

The generated captions and images might not be high quality enough, or across all necessary categories.

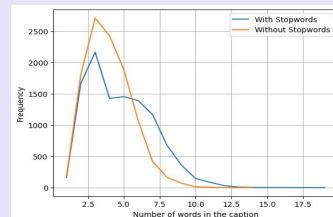


Difficulty metrics

They require very precise word matches and can't capture quality as easily as humans.



Caption distribution



Long tails of rare and unique words.

Limitations of
CNN + RNN

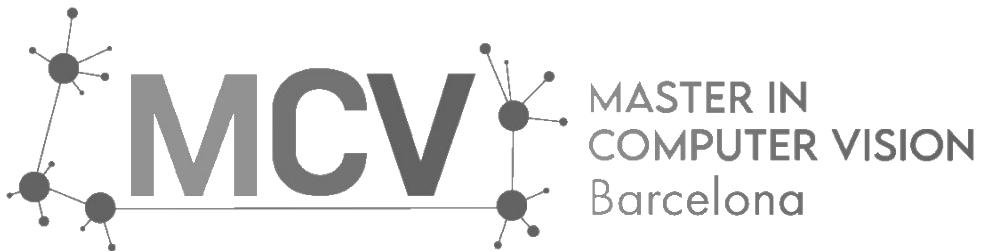
Best **LLM-based**
models often identify
core ingredients

Challenging dataset. Many
terminology

Good prompts in
MLLMs can outperform
fine-tuned models

More data \neq better
results

Challenges in
quantitative evaluation
for **generative models**



Multimodal Recognition Project

Final Session: Project Presentation

Group 6

Module: C5 - Visual Recognition

Coordinators: Ernest Valveny, Carlos Boned, Lei Kang

Abril Piñol

Biel González

Mireia Majó

Alejandro Donaire