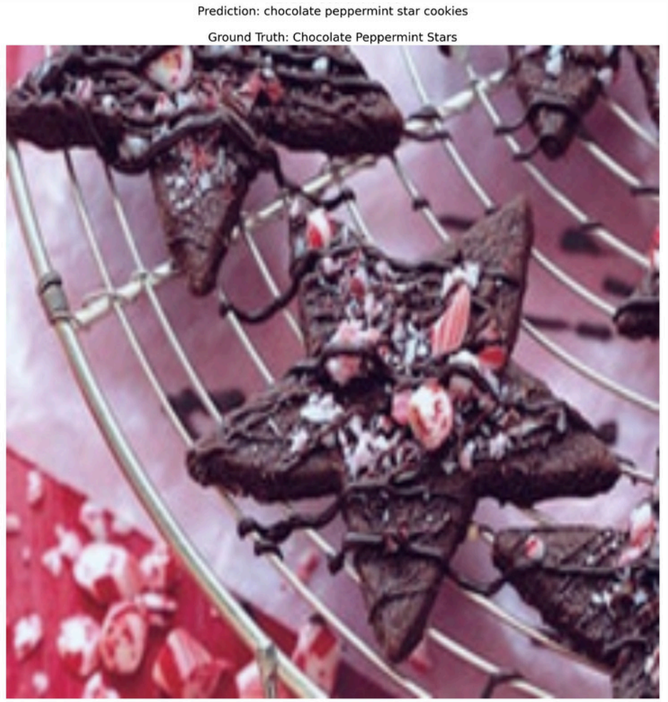
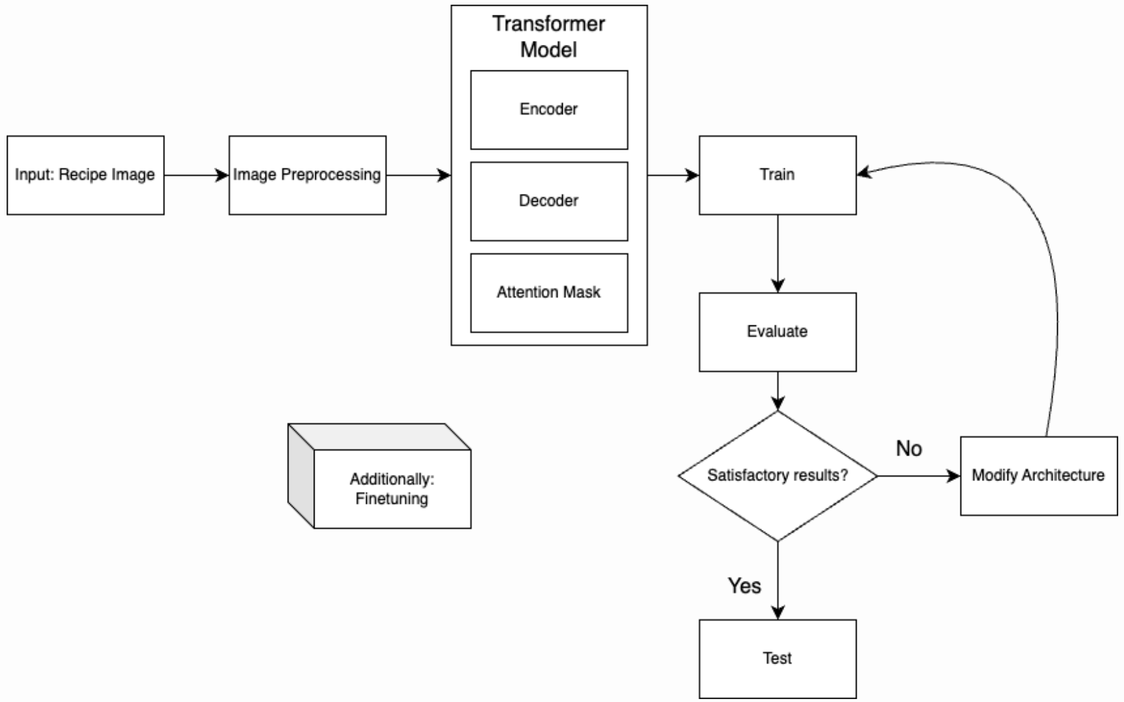
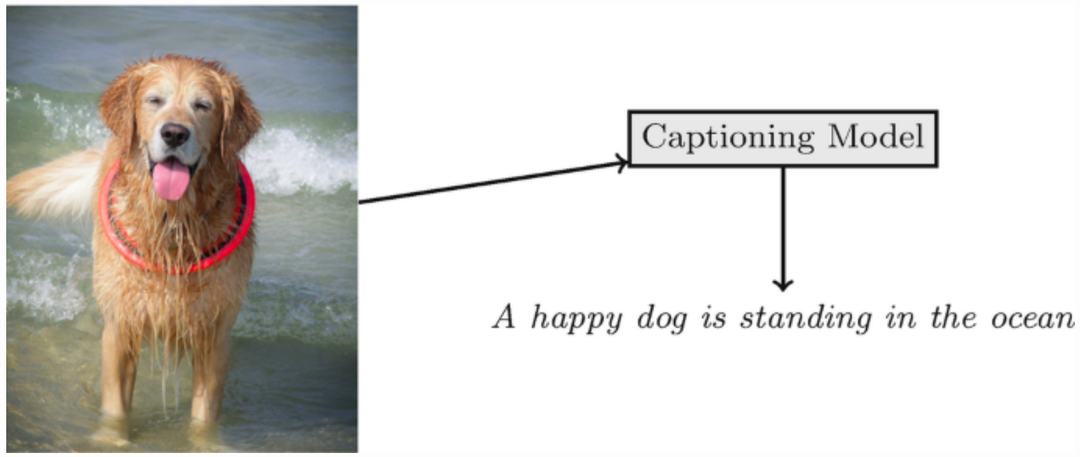


IMAGE CAPTIONING

MARINO OLIVEROS
LUIS DOMENE
ERIC LÓPEZ

GROUP 1





INTRODUCTION

Image captioning generates textual descriptions for images by combining computer vision and natural language processing. Transformers are ideal for this task as they handle sequential data well, capture long-range dependencies, and align image features with text effectively. Focusing on food images and recipes allows us to specialize in culinary terms and create precise captions tailored to this domain, improving relevance for recipe apps and similar applications

We will dive deep in the successes, difficulties and limitations that we have encountered.

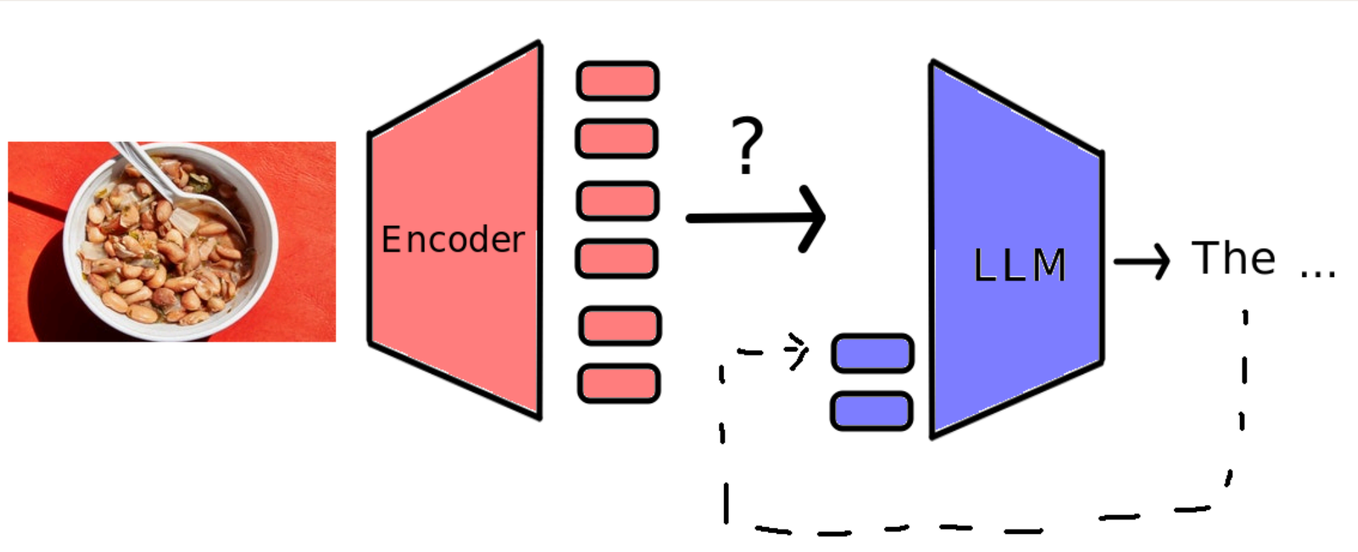
DATASET ANALYSIS

Recipe Dataset: 13,582 images. 13,306 captions. 30 Missing images. 111 Unmatched images. 162 Duplicate captions. 5 Missing captions.

3-ingredient-grilled-orange-margarita.jpg

Image width	Min: 274 Mean: 274.031 Max: 702
Image height	Min: 169 Mean: 169.041 Max: 722
Title length	Min: 3.0 Mean: 32.761 Max: 112.0
Lexicon	Common words in Titles: with, and, Salad, Chicken, Sauce, Grilled… Common words in Ingredients: 1, Cup, Teaspoon, 2, Tablespoons, fresh… Rare words Titles: Board, Thread, Sinigang, Bihon, Eureka… Rare words in Ingredients: Nougat, cup/95, Gluten-free, oolong…

PROPOSED METHOD: VISION ENCODER + LLM

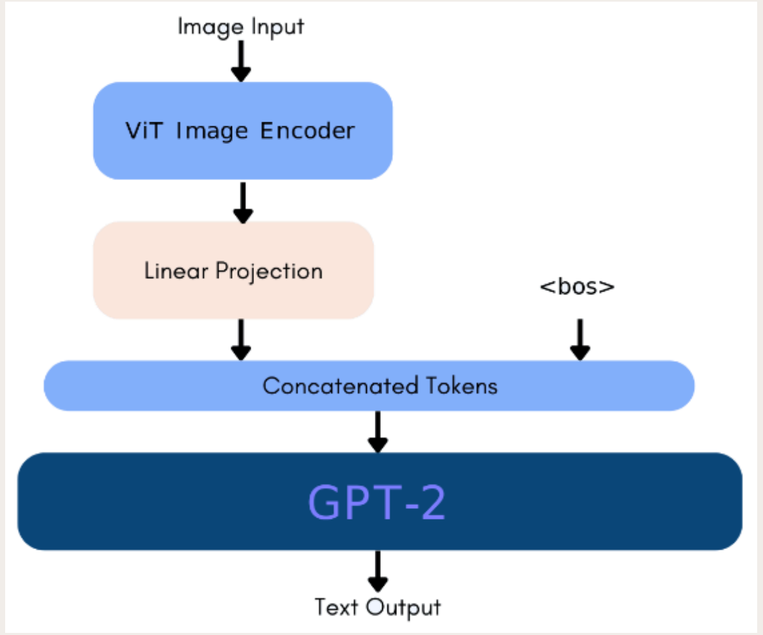


How do we join encoder with LLM?

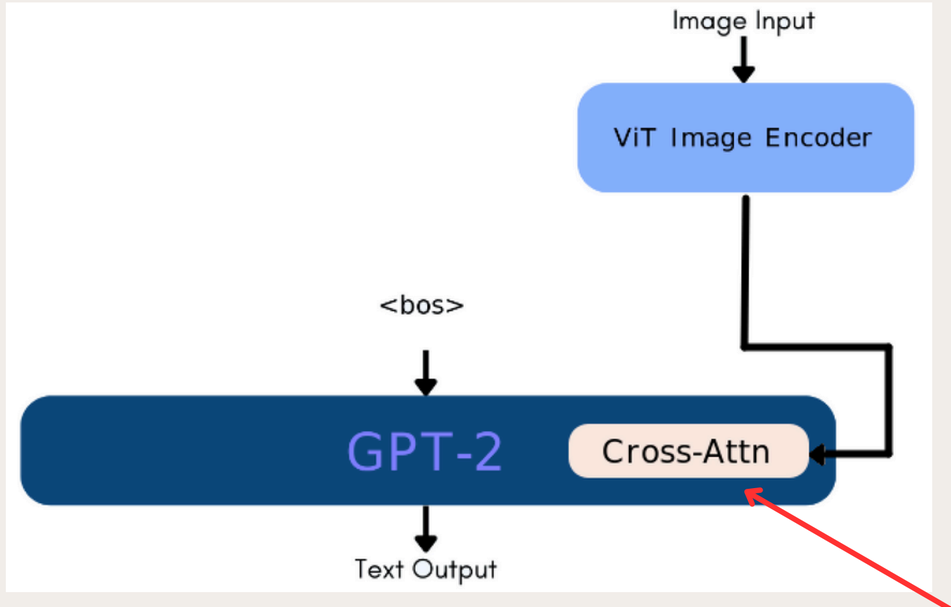
How do we autoregressively generate the caption?

VISION ENCODER DECODER
VIT-GPT2-VED

240M PARAMS



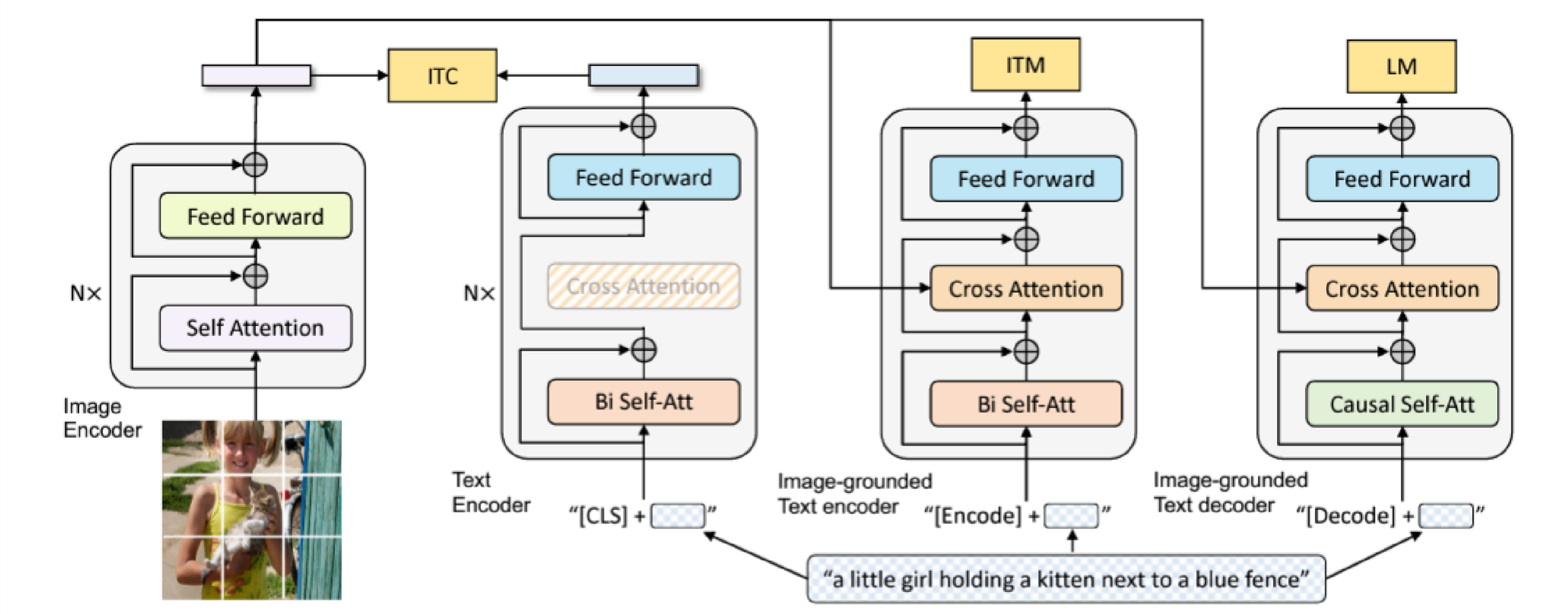
Conventional
Vision Language Model



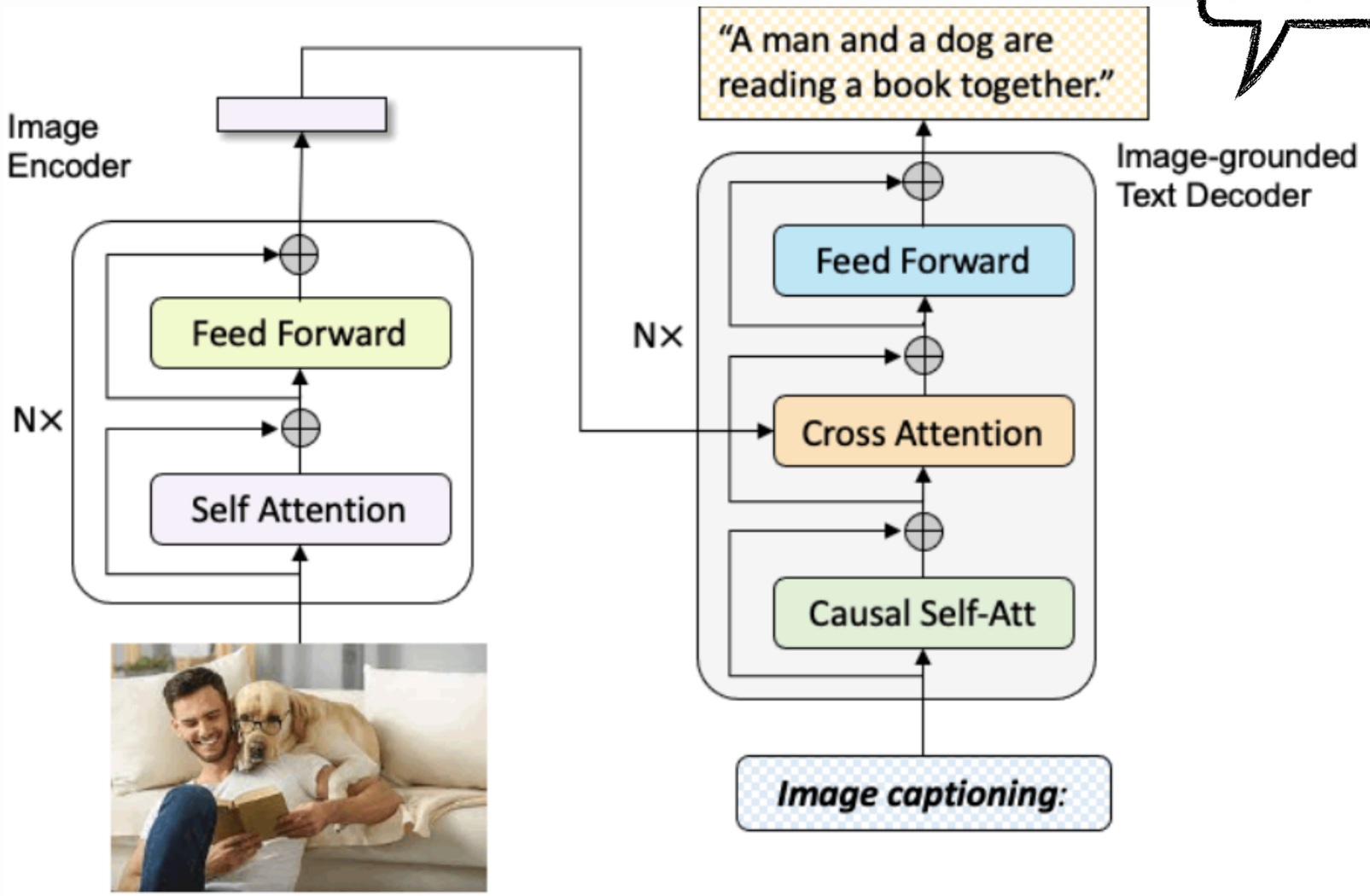
Vision Encoder Decoder (VED)

Train from scratch!

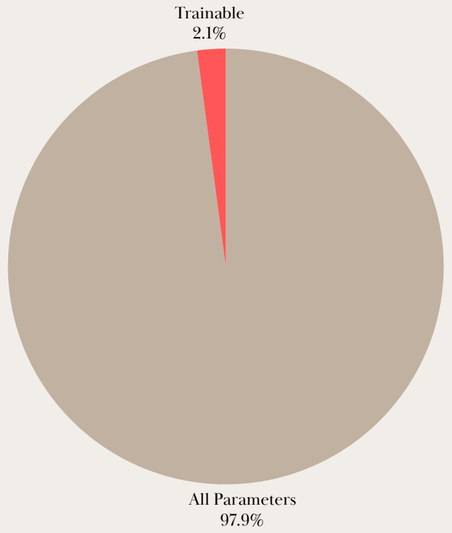
PROPOSED METHOD:
BLIP



252M PARAMS

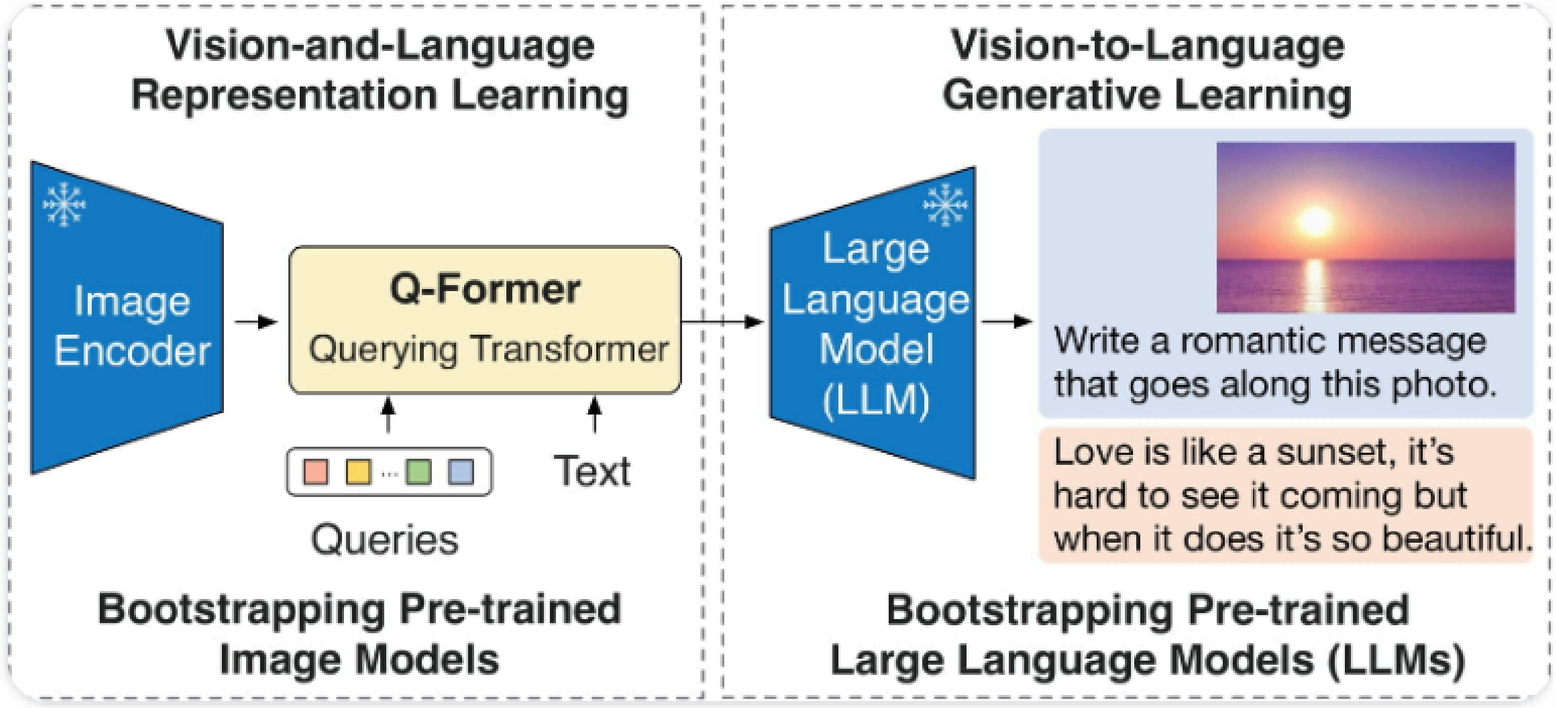


FINETUNNING ATTENTION AND LINEAR LAYERS WITH LORA

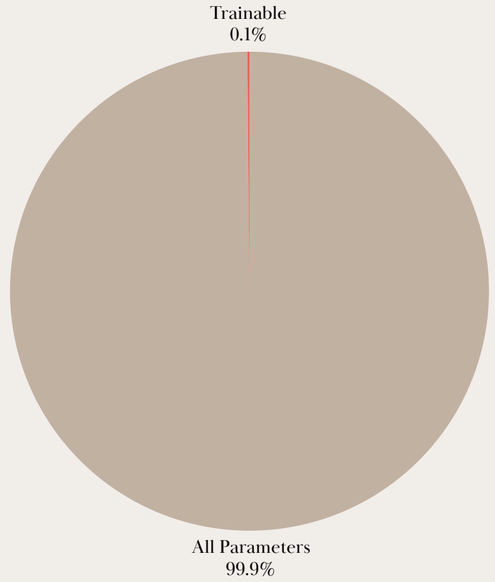


PROPOSED METHOD: BLIP2

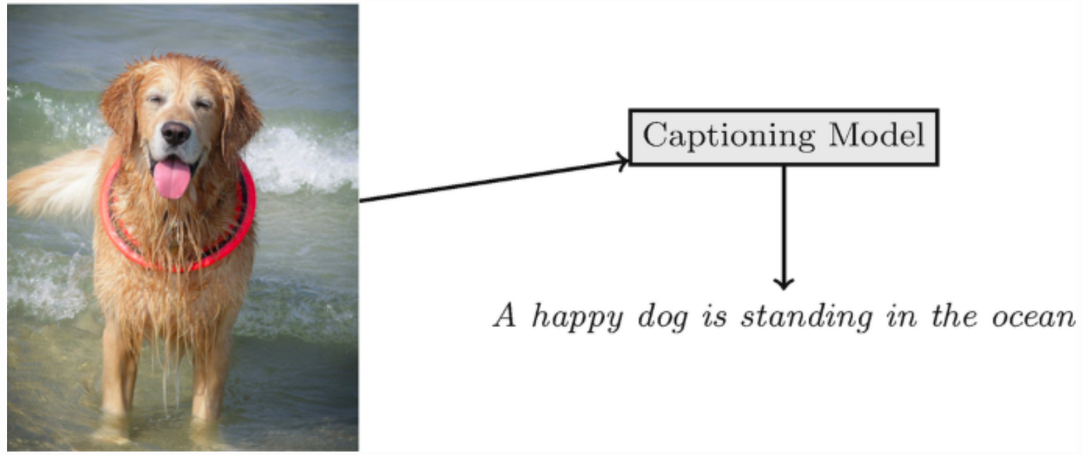
3.7B PARAMS



FINETUNNING WITH LORA



METRICS



OR

A happy pup is in the ocean.

A playful dog is standing in the surf.

A blissful canine is wading by the water.

METRICS

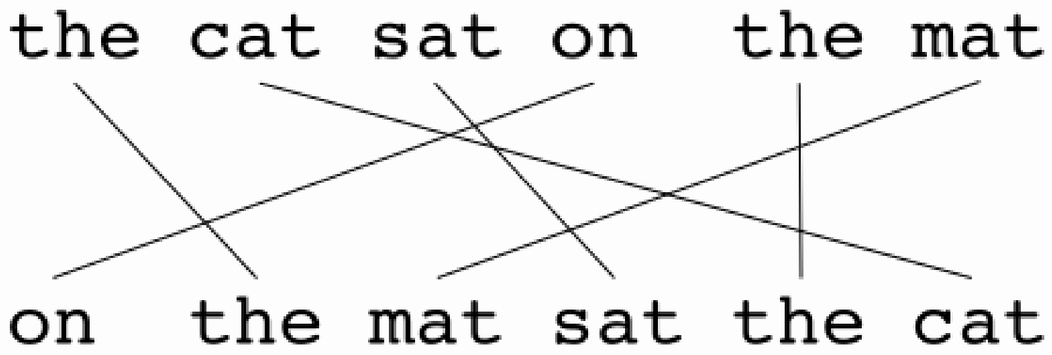
01 **BLEU-1**

02 **BLEU-2**

BLEU is a precision-based metric that evaluates how closely a generated caption matches the reference captions. BLEU calculates n-gram overlap between the predicted captions and reference captions, with BLEU-1, BLEU-2, BLEU-3, and BLEU-4 referring to the overlap of unigrams, bigrams, trigrams, and four-grams, respectively.

$$\text{BLEU} = \min\left(1, \frac{\text{output-length}}{\text{reference-length}}\right)\left(\prod_{i=1}^{4} \text{precision}_i\right)^{\frac{1}{4}}$$

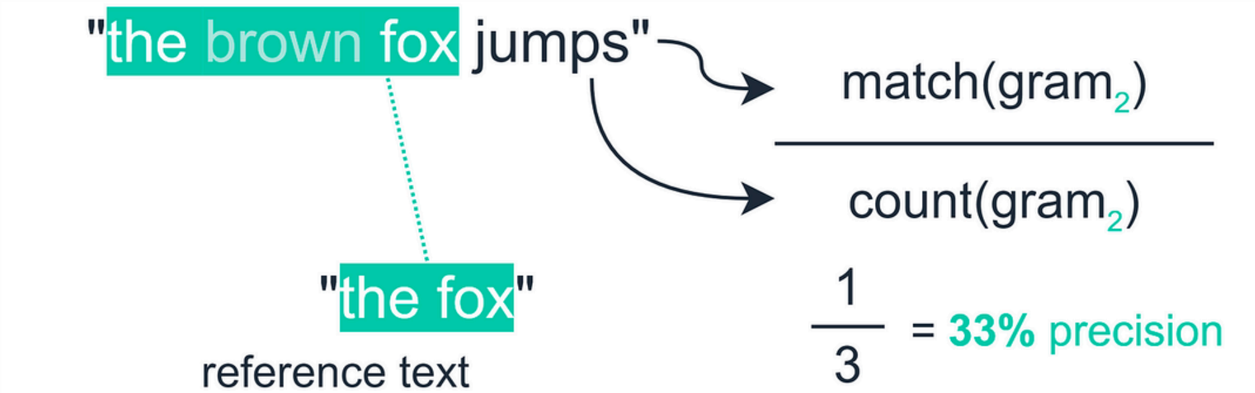
03 METEOR



METEOR is a more comprehensive metric that tries to improve upon BLEU by considering: synonymy, stemming, and word order.

- Weighted average of precision and recall.
- Penalty for word order mismatches + sensitive to caption structure.

04 ROUGE



The **ROUGE** score focuses on recall, measuring the overlap of n-grams, word sequences, or word pairs between the predicted and reference captions.

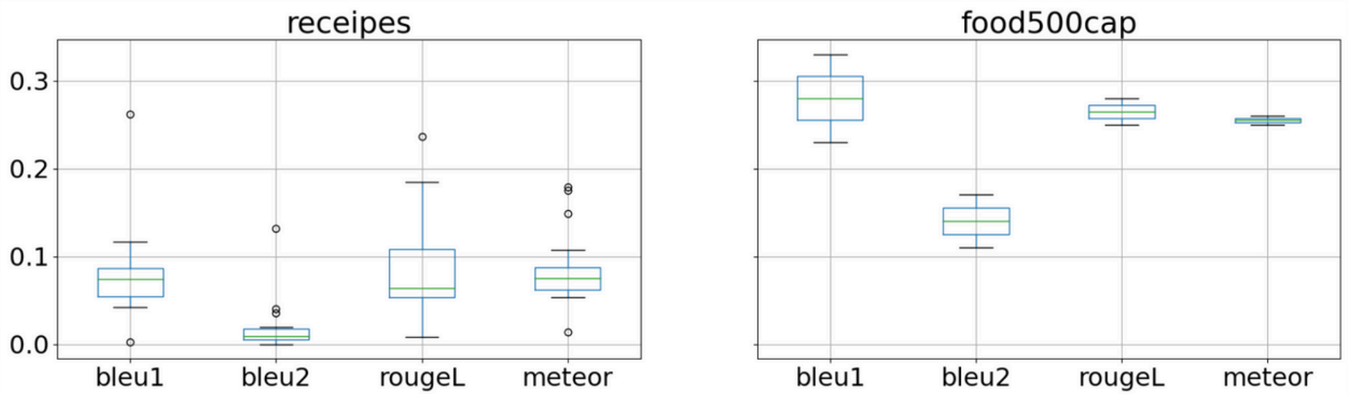
- Good at assessing fluency and syntactic integrity.
- Rewards coherency.

ABLATION STUDIES

Model Name	Use pretrained Encoder	Use pretrained Decoder	Train Encoder	Train Decoder	Contrastive Loss	Inference mode	BLEU-1	BLEU-2	ROUGE-L	METEOR
ViT-GPT2	✓	✓		✓		Greedy	0.075	0.009	0.054	0.083
DINO-GPT2	✓	✓		✓		Greedy	0.058	0.012	0.054	0.070
	✓	✓	✓	✓		Greedy	0.076	0.010	0.031	0.055
	✓	✓	✓	✓	x0.1	Greedy	0.078	0.015	0.062	0.089
	✓	✓	✓	✓	x0.1	Sampling	0.049	0.005	0.054	0.063
	✓	✓	✓	✓	x1	Sampling	0.053	0.000	0.043	0.069
	✓	✓	✓	✓	x10	Sampling	0.042	0.008	0.038	0.057
	✓	✓	✓	✓	x0.1	Beam search	0.096	0.009	0.063	0.081
DINO-SmolLM	✓	✓	✓	✓	x0.1	Beam search	0.003	0.000	0.008	0.014
DINO-GPT2-VED	✓	✓	✓	✓	x0.1	Sampling	0.058	0.000	0.081	0.062
ViT-GPT2-VED	✓	✓	✓	✓	x10	Beam search	0.074	0.016	0.107	0.080
	✓	✓	✓	✓	x10 (clip)*	Beam search	0.089	0.020	0.109	0.084
			✓	✓	x10 (clip)*	Beam search	0.049	0.006	0.064	0.054
		✓	✓	✓	x10 (clip)*	Beam search	0.096	0.018	0.105	0.069
BLIP	✓	✓			x1	Sampling	0.117	0.040	0.134	0.107
	✓	✓	✓	✓	x1	Sampling	**0.262**	**0.132**	**0.236**	**0.179**
BLIP2	✓	✓			x1	Sampling	0.078	0.003	0.185	0.175
	✓	✓	✓	✓	x1	Sampling	0.074	0.036	0.131	0.149

ABLATION STUDIES

Trying a different dataset: Food500Cap



EXAMPLES OF FAILURES

VIT-GPT2-VED: “Chocolate-Pistachio Short Ribs with Pistachios and Fennel-C”
BLIP: “arugula salad with mushrooms and parmesan”

GT: “Sauteed Dandelion Greens”

VIT-GPT2-VED: “Grilled Chicken Thighs with Fennel, Peas, and Cucumber Salad (”
BLIP: “pecan - caramel tart”

GT: “Bittersweet Chocolate Pecan Pie”

VIT-GPT2-VED: “Coconut-Pistachio Popsicles with Lemon-Ginger Dressing Sauce and ”
BLIP: “salted butter cookies”

GT: “Sage-Scented Shortbread”

EXAMPLES OF SUCCESSES

Receipes

BLIP: “tomato and corn pie”

GT: “tomato and corn pie”

BLIP: “spinach and mint soup”

GT: “spinach and mint soup”

Food500Cap

BLIP: “a muffin with a bite of strawberry jam on top of it, and a few muffins on the bottom”

GT: “red juicy strawberry jam was stuffed into baked golden cakes.”

BLIP: “a set of rice balls with red pepper on top, served with a bowl of soup and a small bowl of soup beside it”

GT: “four steamed meatballs covered with glutinous rice on the surface topped with a goji and some cilantro.”

REAL LIFE DISH EXAMPLE

BLIP: baked fish with potatoes and carrots
BLIP2: Fish with Tomaotes, Parsley, and Garlic Sauce, and Tomatoes and Parsley
BLIP2+Prompt: cooked in a cast-iron pan with tomatoes, potatoes and onions

BLIP: Pasta with meat and Mushrooms
BLIP2: Pasta with Lamb and Prawns (Pasta con le Pane) with Pecorino
BLIP2+Prompt: Starches, the beans, the lentils, the chicko, the onions, the garlic

BLIP: Pasta with bacon and parmesan
BLIP2: Chili Noodles with Sp (Noodles)(Noodles)
BLIP2+Prompt: saucepapollo, a traditional italian sauce, with dried tomatoes, onions and spices

CONCLUSION

- DATASET COMPLEXITY
- THE COLLABORATION BETWEEN DIFFERENT MODULES IS NOT AS STRAIGHTFORWARD AS WE THOUGHT.
- METRICS LIKE BLEU MIGHT NOT FULLY CAPTURE QUALITATIVE IMPROVEMENTS

FUTURE PLAN

- TRAIN A CUSTOM TOKENIZER TO LEARN SPECIFIC VOCABULARY
- FULL FINETUNNING OF BLIP
- FURHTER RESEARCH IN PROMPT ENGINEERING AS SEEN WITH THE MODIFICATION OF BLIP2