

FUSECAP: Leveraging Large Language Models for Enriched Fused Image Captions

Noam Rotstein, David Bensaïd, Shaked Brody, Roy Ganz, Ron Kimmel
Technion - Israel Institute of Technology

Citations: 119
Accepted to WACV 2024

Presented by Group 2: Mahad Ali, Anthony Jackson, Rafeeq Shodeinde, Isaac Tuckey

Presentation outline

- Background and Motivation
- Caption Fusion
 - Leveraging visual experts
 - LLM fuser
 - Dataset generation
- Evaluation
 - Qualitative Evaluation
 - FuseCap-trained model performance

Motivation

- Captioning is a valuable capability of VLMs
- Capability to caption is largely dependant on data used
 - Large datasets reliant on scraping the web
 - Resultant model captions overlook details
- Generating high quality captions by hand, at scale, isn't feasible

Contributions

- FuseCap: a Framework for improving caption quality
- A large scale dataset of image-text pairs
 - 12 million pairs from assorted other datasets
 - Captions have been enriched using FuseCap

Visualization



GIT: a red motorcycle parked on a road near a beach

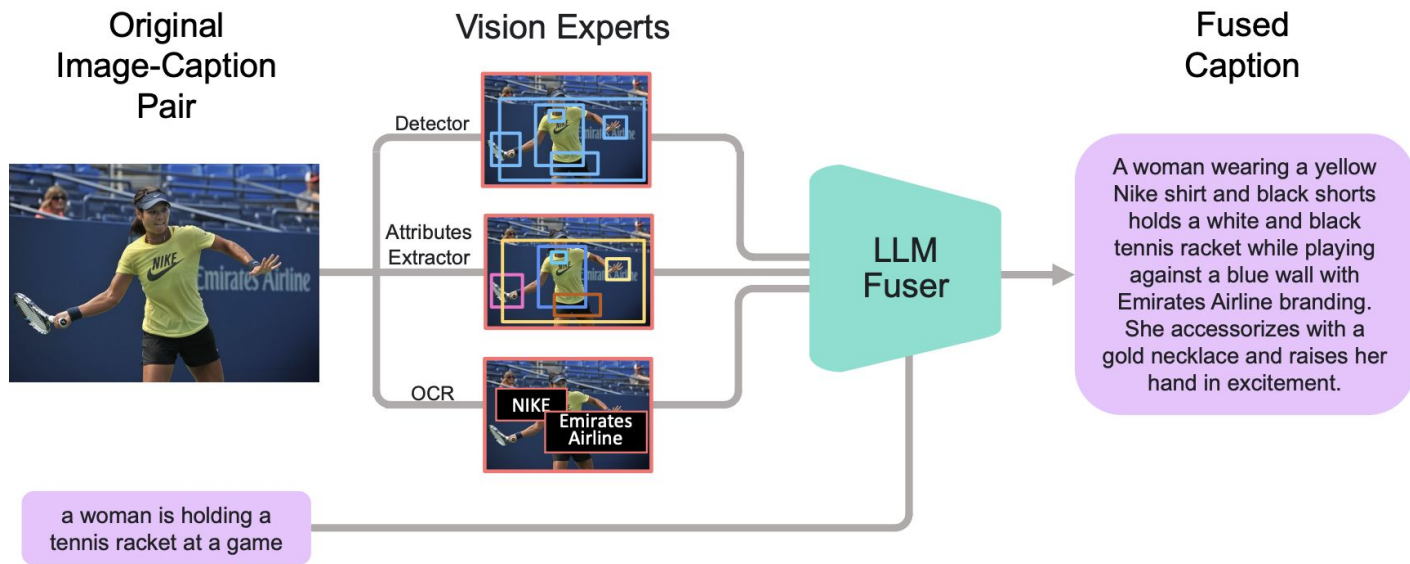
OFA: motorcycle parked on the beach

Prismer: motorcycle parked on the beach

BLIP2: a red motorcycle parked in a parking lot next to a fence

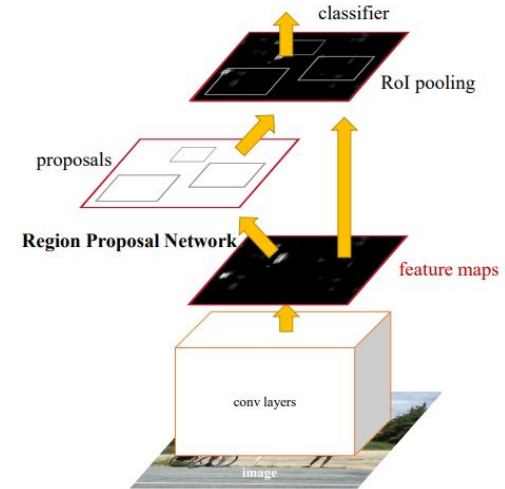
Ours: a red motorcycle with a leather and black seat is parked on the side of the road, surrounded by a wood fence and tall palm trees the clear blue sky provides a serene backdrop

FUSECAP



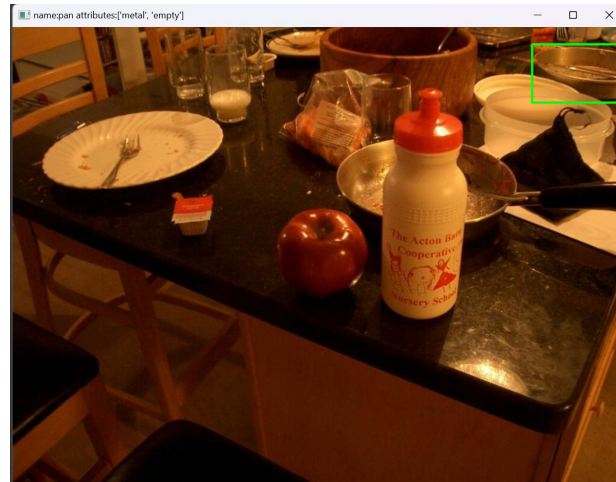
Object Detector

- Faster-RCNN with ResNeXt-152 backbone
 - Pretrained on multiple detection datasets
 - Fine-tuned on Visual Genome
 - Knowledge on 1.6k classes
- Threshold of .7 applied to bounding boxes



Attribute Detector

- Faster-RCNN with ResNeXt-152 Backbone
- Attribute Classifier added to pretrained object detector
 - Descriptors for objects in the image
 - EX: Colors, size (small, large, etc), material (steel, wooden, etc)
- Detector fine-tuned using Visual Genome
 - 400 of the 2.8 million possible attributes
- Threshold of .2 used for attribute predictions

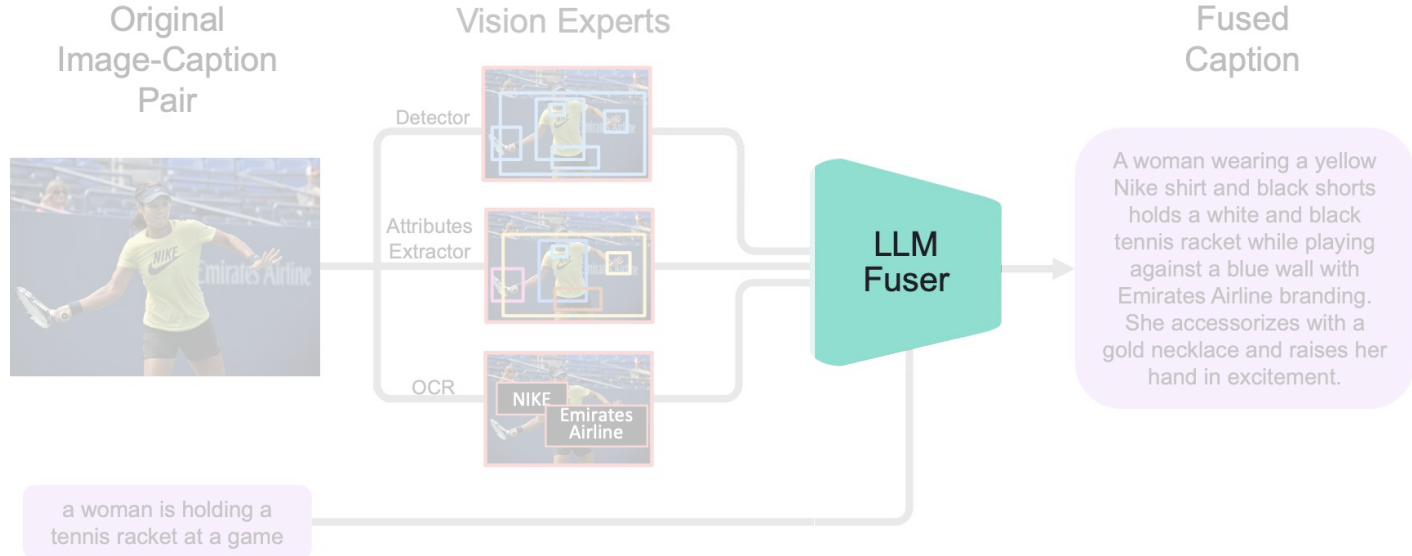


Text Detection Module

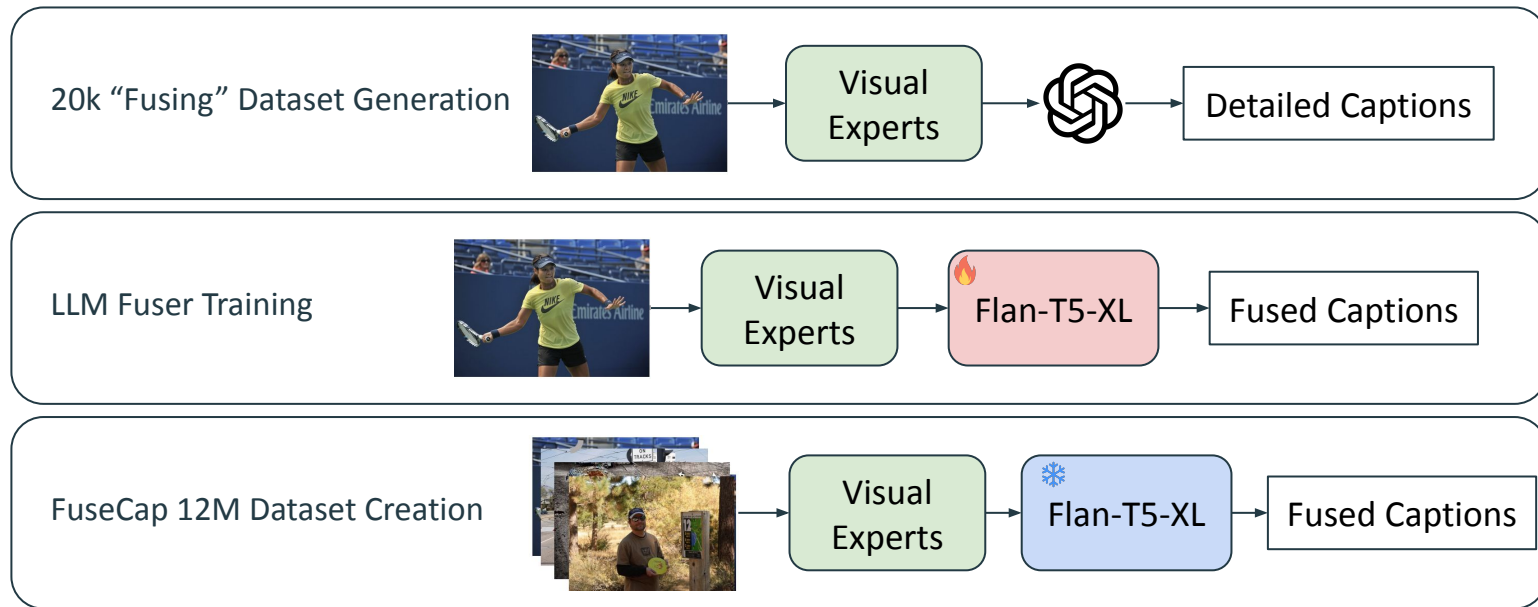
- Character Region Awareness for Text Detection (CRAFT)
 - CNN architecture
 - Produces bounding boxes for words or characters
- Scene Text Recognition with Permuted Autoregressive Sequence Models (Parseq)
 - Encoder-Decoder Architecture for OCR
 - ViT used to encode CRAFT bounding box contents
- Unique assignment of text to objects
 - Uses bounding boxes which contain text bounding box
 - assigns to object w/ smallest bounding box



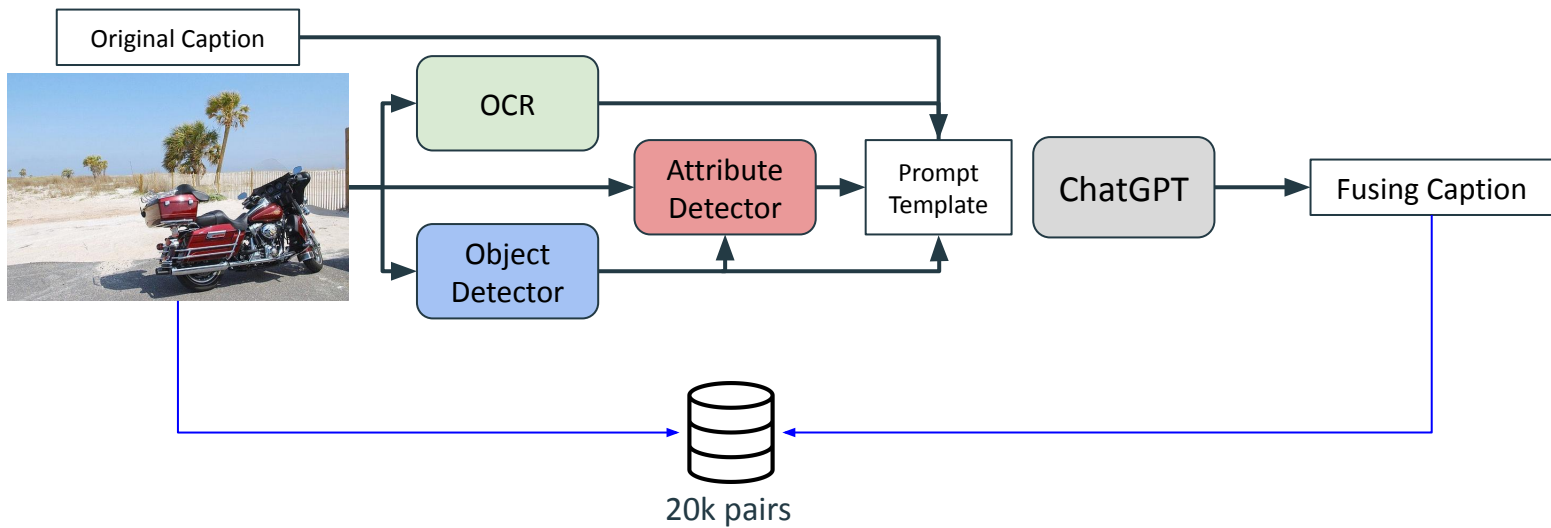
LLM Fuser



LLM Fuser Training



Fusing Caption Dataset Creation

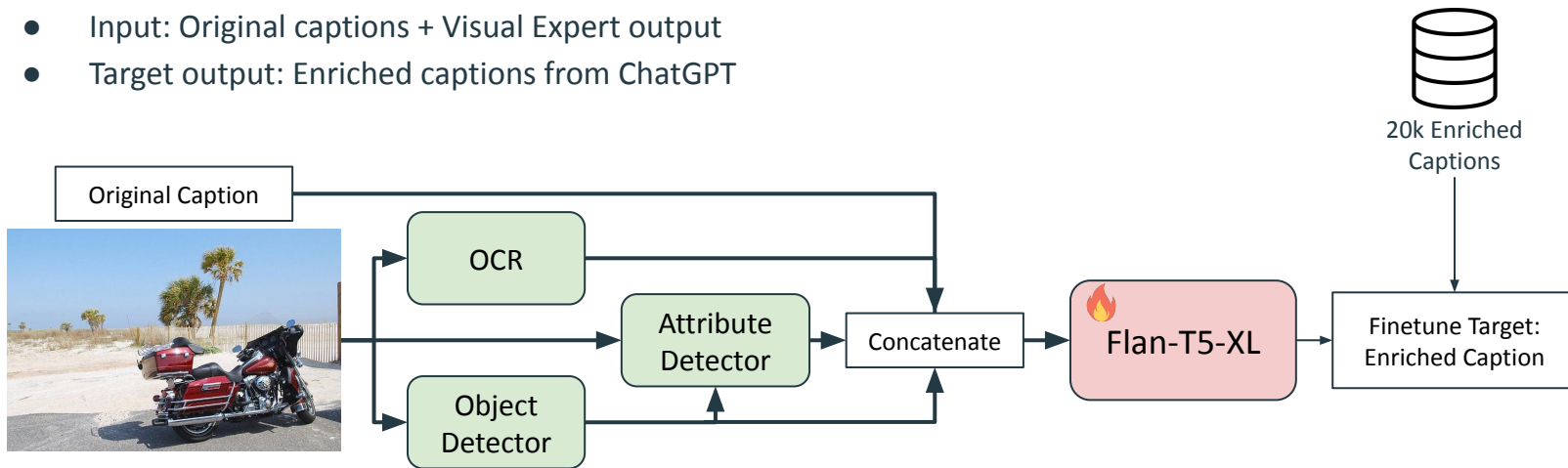


ChatGPT Template

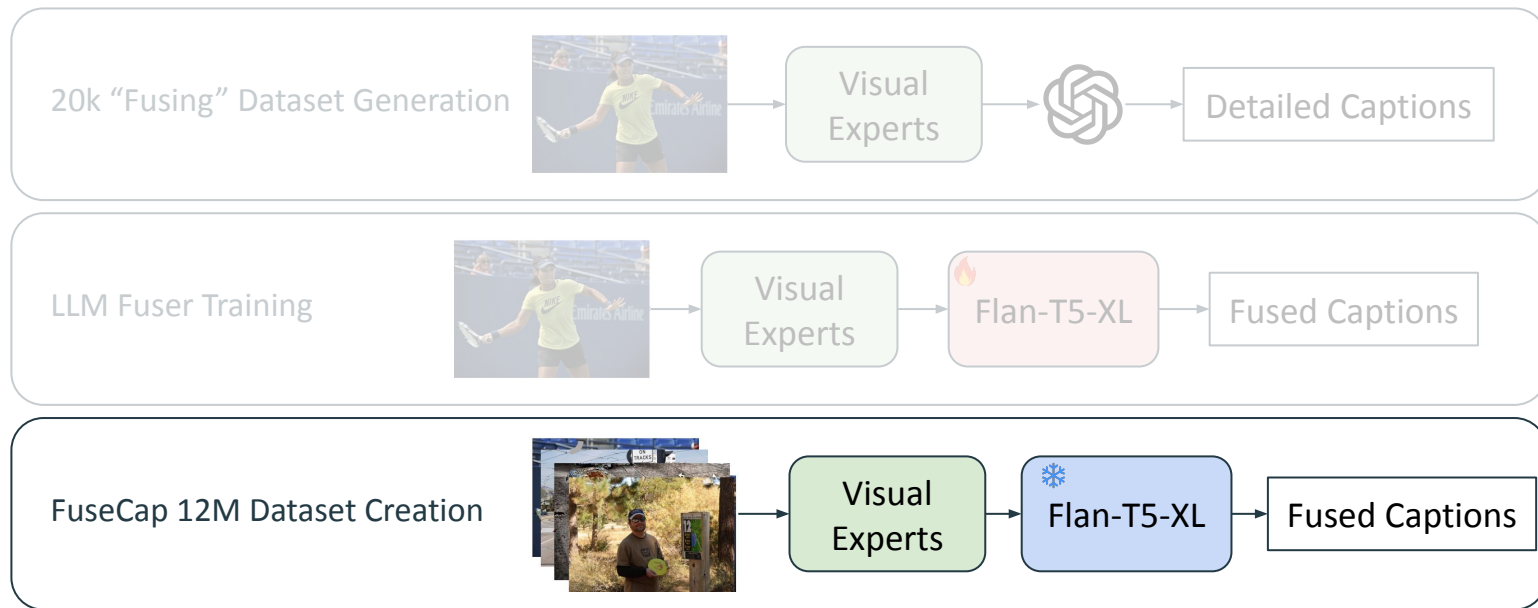
”A caption of an image is given: *original caption*.
The following objects are detected in the image from left to right:
A a_1^1, \dots, a_1^{k-1} and a_1^k o_1 [with the following text: t_1].
:
A $a_n^1, \dots, a_n^{k_n-1}$ and $a_n^{k_n}$ o_n [with the following text: t_n].
Write a comprehensive and concise caption of the scene using the objects detected.”

LLM Fine-tuning

- Flan-T5-XL Checkpoint
- Input: Original captions + Visual Expert output
- Target output: Enriched captions from ChatGPT



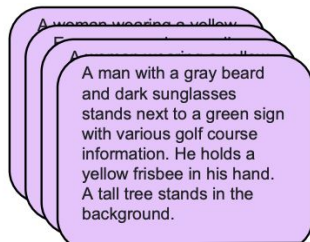
FuseCap Dataset



BLIP Captioning Model

- FUSECAP dataset used to optimize ITC, ITM, and LM Loss in BLIP
- Fine tuned on the COCO dataset with the LM loss
- Context length increased from 30 -> 60 tokens to improve comprehensive caption generation

Enriched
Image-Caption
Pair



Caption
Generation
Pretrain+Finetune



Experiments

Qualitative Evaluation

- Human study
- *“Does caption 2 provide an additional meaningful and truthful description of the image compared to caption 1?”*



Original: Mhmm, some clouds in the sky

Ours: A woman wearing dark sunglasses stands next to a red car with a black license plate reading 166882, PRI. The car has off and round headlights, a chrome and silver bumper, a black tire, and a red door. The cloudy and white sky is visible in the background.

CLIPScore comparison

- CLIPScore: Cosine similarity between image and text features
 - Mean: mean score
 - Voting: choose between captions

Dataset	Captions	Mean	Voting
COCO	Original	76.7	31.7%
	FUSECAP	80.3	67.6%
SBU	Original	71.9	32.1%
	FUSECAP	75.5	60.2%
CC	Original	72.6	34.7%
	FUSECAP	75.4	59.7%

Image-Text Retrieval Task

Model	COCO Retrieval					
	img \rightarrow text			text \rightarrow img		
	R@1	R@5	R@10	R@1	R@5	R@10
BLIP \dagger	75.1	92.7	96.4	58.2	82.4	89.2
BLIP-L	82.4	95.4	97.9	65.2	86.3	91.8
BLIP2	85.4	97.0	98.5	68.3	87.7	92.6
BLIP * _{FUSECAP}	97.2	99.5	99.9	93.0	97.4	98.3

ITR with Generated Captions

- Captions generated by BLIP models

							COCO Retrieval											
							img → text			text → img								
Model	R@1 R@5 R@10			R@1 R@5 R@10			Model	R@1 R@5 R@10			R@1 R@5 R@10							
BLIP†	75.1	92.7	96.4	58.2	82.4	89.2	BLIP†	56.3	83.0	90.3	54.5	81.2	88.7					
								-18.8%	-9.7%	-6.1%	-3.7%	-1.2%	-0.5%					

Image Captioning

- Metric: CLIPScore

Model	Images	Parameters	Val	Test
BLIP†	12M	247M	75.2	75.3
BLIP-L	129M	470M	76.1	76.0
OFA	20M	470M	76.6	76.4
GIT	800M	700M	77.1	77.0
BLIP2-G-OPT _{2.7}	129M	3.8B	77.8	77.5
Prismer	13M	1.6B	76.7	76.7



GIT: a man riding a horse with a dog in the background.

OFA: a man riding on the back of a white horse

Prismer: A man riding a horse next to a small dog.

BLIP2: a man riding a horse with a dog in the field

Ours: a man wearing a red hat and blue jeans rides a white horse with a long tail, while a small white dog follows closely behind



GIT: a man riding a small motorcycle in a parking lot.

OFA: a man riding a motorcycle in a parking lot

Prismer: A man riding a motorcycle in a parking lot.

BLIP2: a man riding a motorcycle in a parking lot with tents

Ours: a man wearing a white shirt and blue jeans rides a motorbike in a parking lot surrounded by white and yellow tents, with a white line marking the edge of the parking

Large-Scale Dataset Influence

Pre-training Data	Fine-tune +Test Data	B@4	CIDEr	SPICE
Standard	Standard	37.8	126.5	22.9
FUSECAP	Standard	38.4	128.7	23.0
Standard	FUSECAP	35.4	111.4	25.0
FUSECAP	FUSECAP	37.3	123.1	26.8

Conclusion

- FUSECAP utilizes visual experts to extract meaningful information from images.
- LLM fuses the data into the existing captions to yield better captions.