Shanghai Jiao Tong University

上海人工智能实验室 Shanghai Artificial Intelligence Laboratory

Zero-shot Composed Text-Image Retrieval

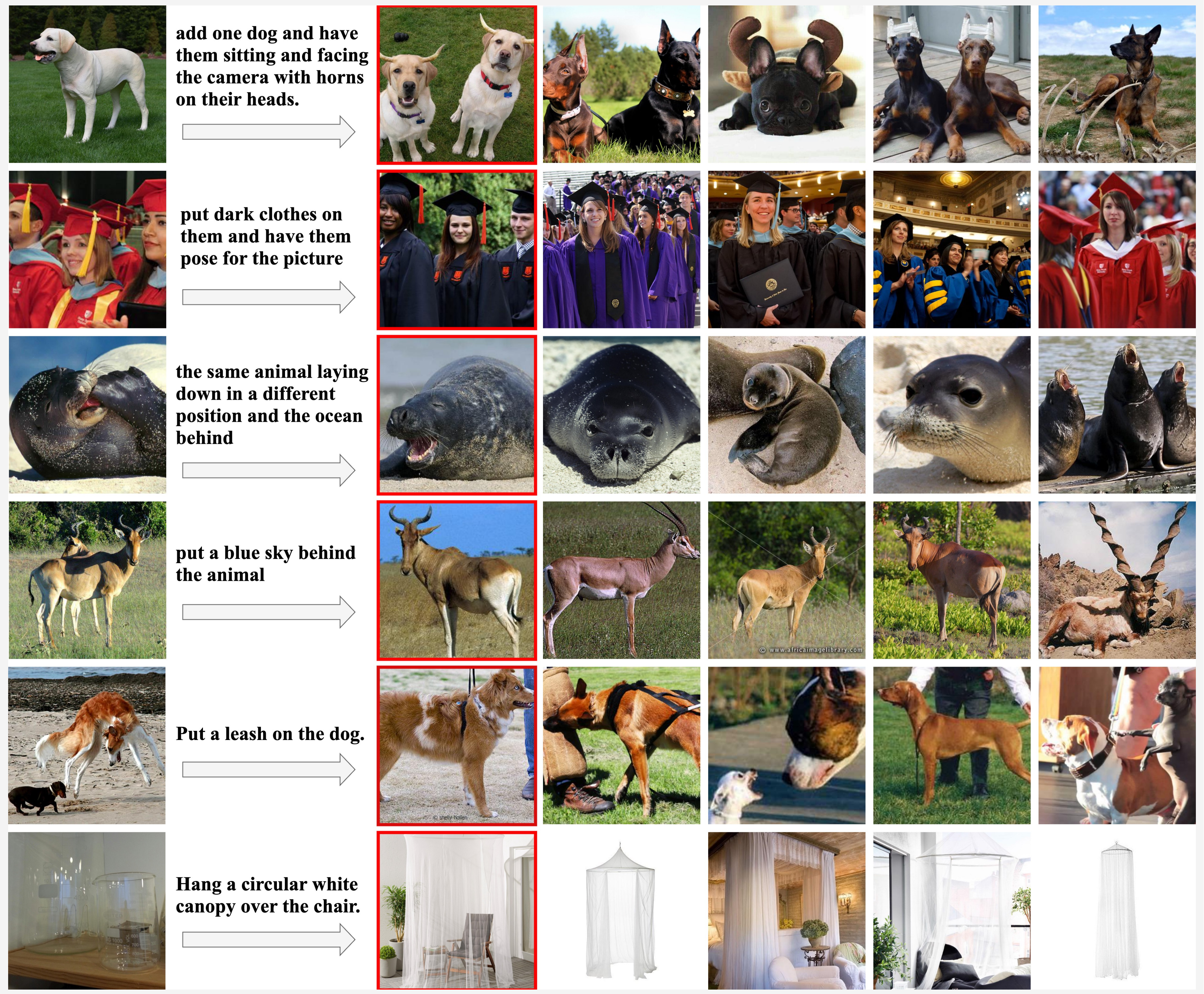
Yikun Liu[1,2], Jiangchao Yao[1,3], Ya Zhang[1,3], Yanfeng Wang[1,3], Weidi Xie[1,3]

[1]CMIC, Shanghai Jiao Tong Uniersity [2]Beijing University of Posts and Telecommunications [3]Shanghai AI Laboratory

BMVC 2023

Problem Scenario

- Composed image retrieval (CIR) aims to train a model that can fuse multi-modal information to accurately retrieve images that match the query.

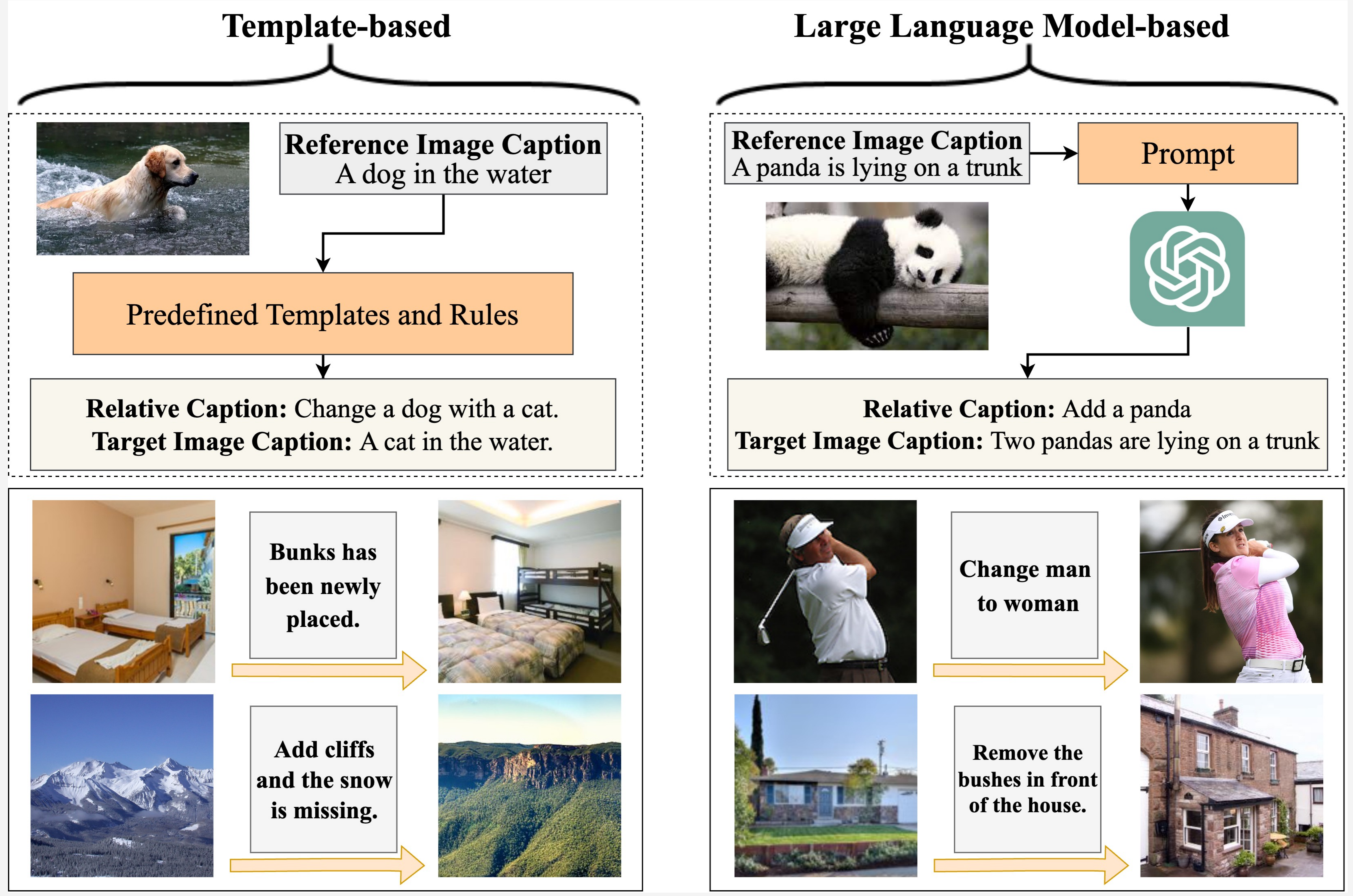

add one dog and have them sitting and facing the camera with horns on their heads.
put dark clothes on them and have them pose for the picture
the same animal laying down in a different position and the ocean behind
put a blue sky behind the animal
Put a leash on the dog.
Hang a circular white canopy over the chair.

Our Contribution

- **Datasets:** we initiate a scalable pipeline to automatically construct datasets for training CIR model, by simply exploiting a large-scale dataset of image-text pairs.
- **Architecture:** we introduce TransAgg, which employs a simple yet efficient fusion mechanism, to adaptively combine information from diverse modalities;
- **Results:** our proposed approach either performs on par with or significantly outperforms the existing state-of-the-art (SOTA) models.

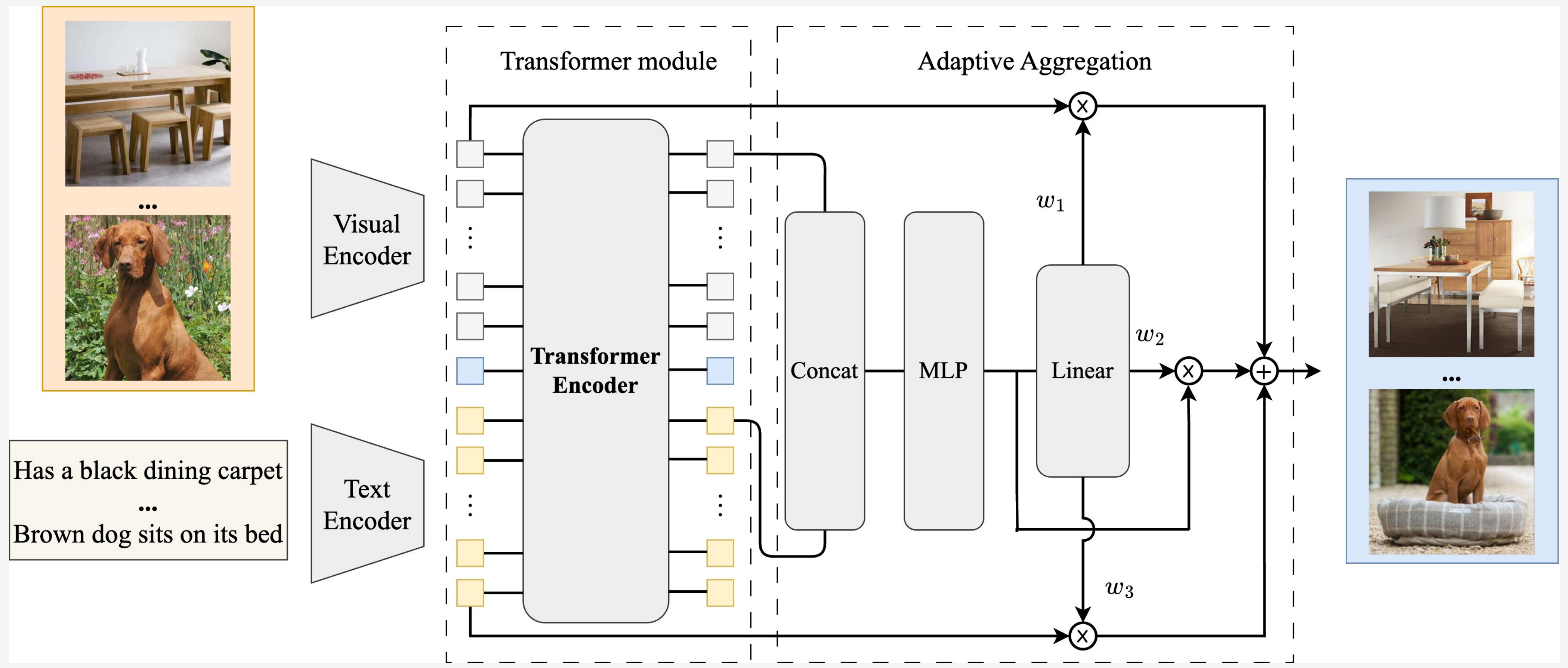
Project Page *Paper* *Code*

The Dataset Construction Procedure


Template-based
Large Language Model-based
Reference Image Caption
A dog in the water
Predefined Templates and Rules
Relative Caption: Change a dog with a cat.
Target Image Caption: A cat in the water.
Reference Image Caption
A panda is lying on a trunk
Prompt
Relative Caption: Add a panda
Target Image Caption: Two pandas are lying on a trunk
Bunks has been newly placed.
Add cliffs and the snow is missing.
Change man to woman
Remove the bushes in front of the house.

- For one image-caption sample, we can revise its caption and use the resulting edited caption as a query
- Retrieve the target image with similar caption, where we adopt Sentence Transformer.
- We obtain different training datasets depending on the different approaches for revising captions.

TransAgg (Architecture)



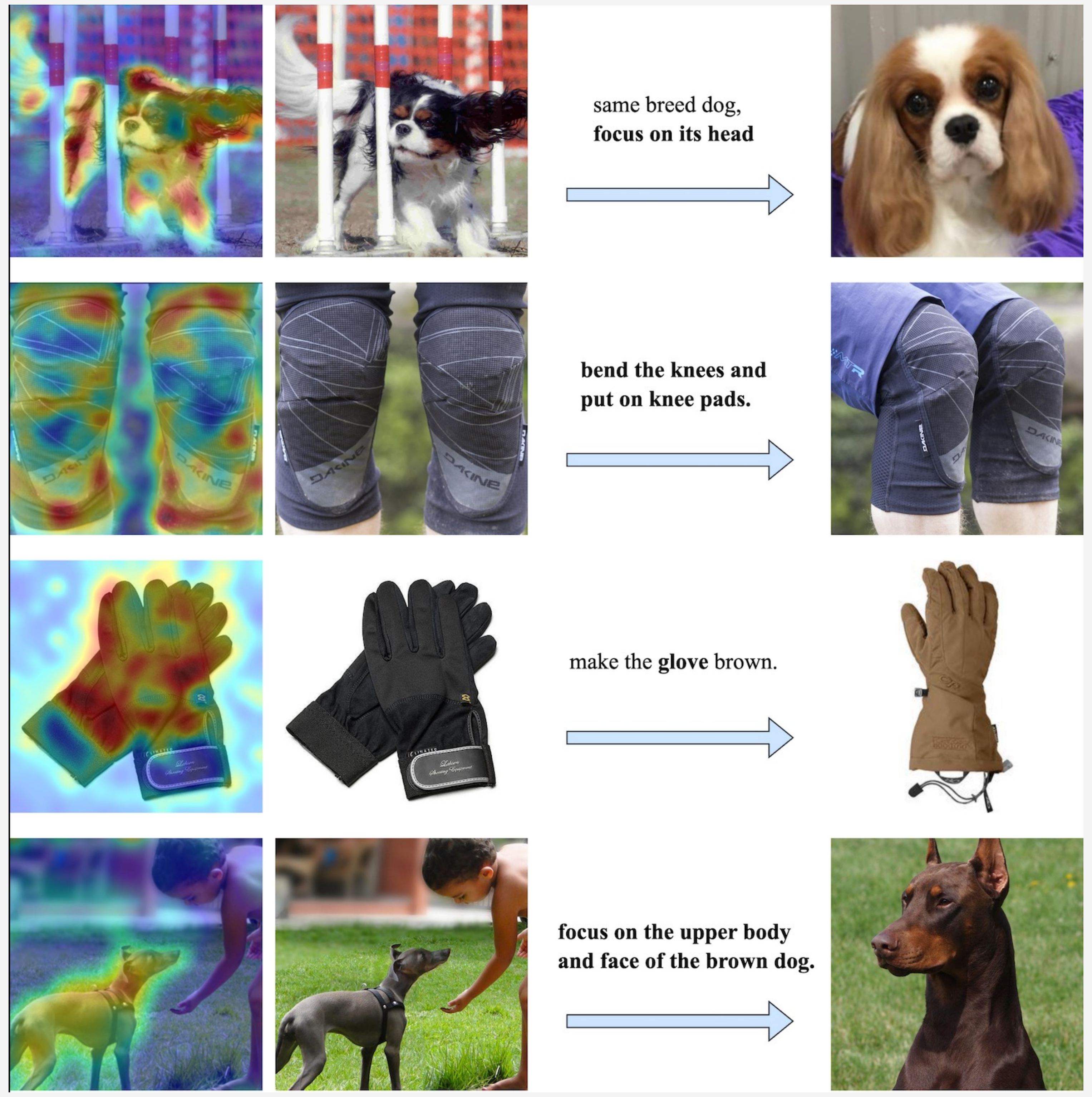
- Encoders to extract features from visual and textual inputs respectively;
- A Transformer module to capture the interaction between two modalities;
- An adaptive aggregation module that combats modal redundancy and fuses the features together.

Experiment Results

Method	Zero-shot	# Triplets	CIRR R@1	CIRR R@5	CIRR R@50	CIRR R_{Subset}@1	FashionIQ R@10	FashionIQ R@50	FashionIQ Average
Pic2Word [25]CVPR'2023	✔	-	23.90	51.70	87.80	-	24.70	43.70	34.20
PALAVRA [5]ECCV'2022	✔	-	16.62	43.49	83.95	41.61	19.76	37.25	28.51
SEARLE-XL-OTI [2]arXiv'2023	✔	-	24.87	52.31	88.58	53.80	27.61	47.90	37.76
CompoDiff w/T5-XL [9]arXiv'2023	✔	18m	19.37	53.81	90.85	28.96	**37.36**	50.85	44.11
CASE Pre-LaSCo.Ca. [15]arXiv'2023	✔	360k	35.40	65.78	**94.63**	64.29	-	-	-
TransAgg (Laion-CIR-Template)	✔	16k	**38.10**	68.42	93.51	**70.34**	32.07	53.26	42.67
TransAgg (Laion-CIR-LLM)	✔	16k	36.71	67.83	93.86	66.03	32.77	53.44	43.11
TransAgg (Laion-CIR-Combined)	✔	32k	37.87	**68.88**	93.86	69.79	34.36	**55.13**	**44.75**
CLRPLANT w/OSCAR [18]ICCV'2021	✘	-	19.55	52.55	92.38	39.20	18.87	41.53	30.20
ARTEMIS [6]ICLR'2022	✘	-	16.96	46.10	87.73	39.99	26.05	50.29	38.17
CLIP4CIR [1]CVPRW'2022	✘	-	38.53	69.98	95.93	68.19	38.32	61.74	50.03
BLIP4CIR+Bi [19]arXiv'2023	✘	-	40.15	73.08	96.27	72.10	43.49	67.31	55.40
CASE [15]arXiv'2023	✘	-	48.00	79.11	97.57	75.88	48.79	70.68	59.74

- On CIRR dataset, our proposed model achieves state-of-the-art results in all metrics except for Recall@50;
- On FashionIQ dataset, our proposed TransAgg model trained on the automatically constructed dataset also falls among the top2 best models.

Explainability Heatmaps


same breed dog, focus on its head
bend the knees and put on knee pads.
make the glove brown.
focus on the upper body and face of the brown dog.