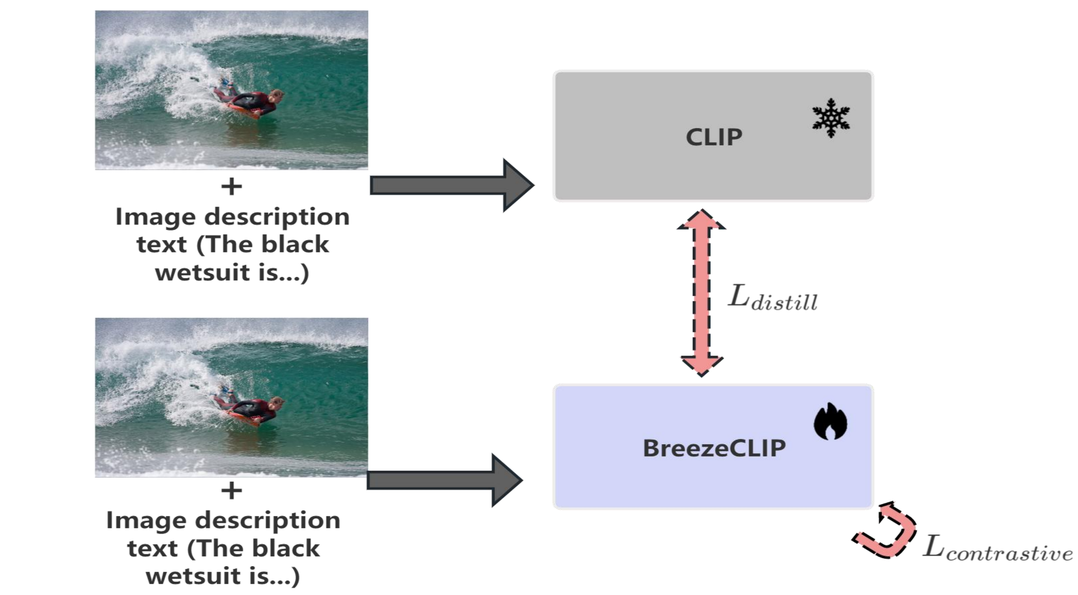
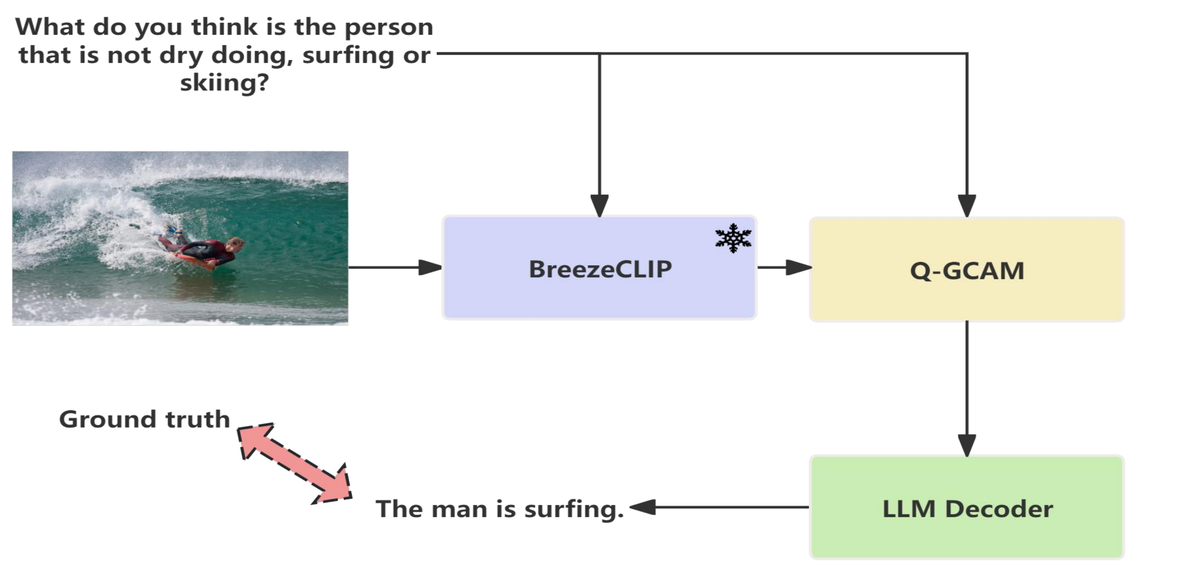
BcQLM: Efficient Vision-Language Understanding with Distilled Q-Gated Cross-Modal Fusion

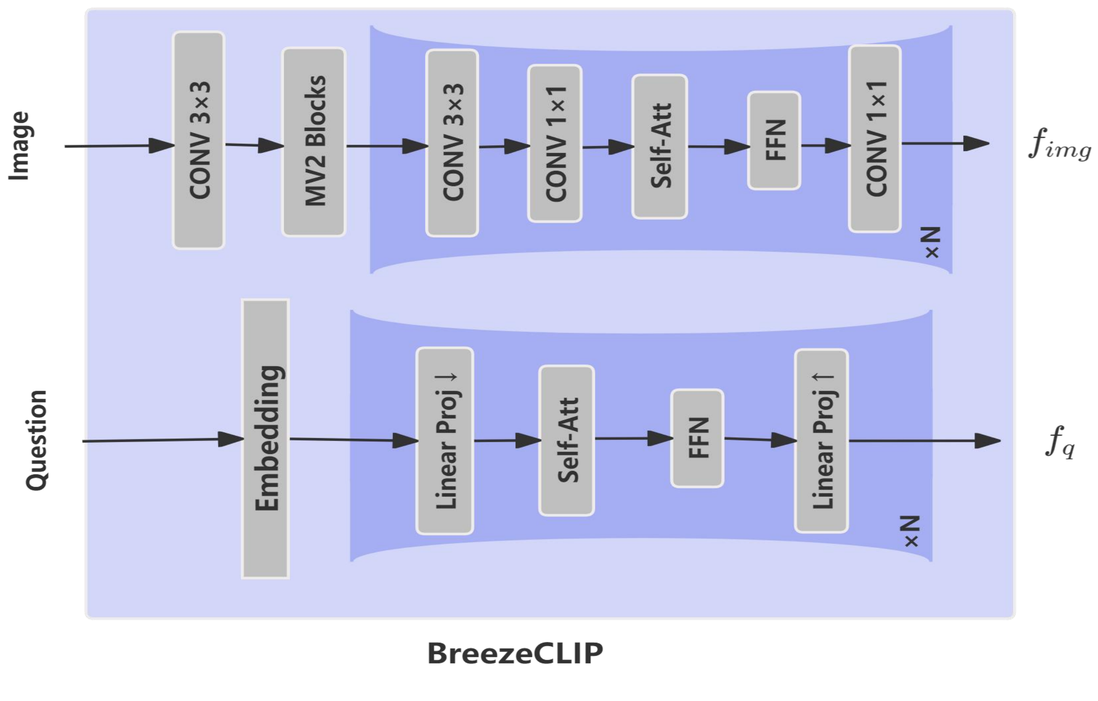
Sike Xiang and Shuang Chen and Amir Atapour-Abarghouei*
Department of Computer Science, Durham University

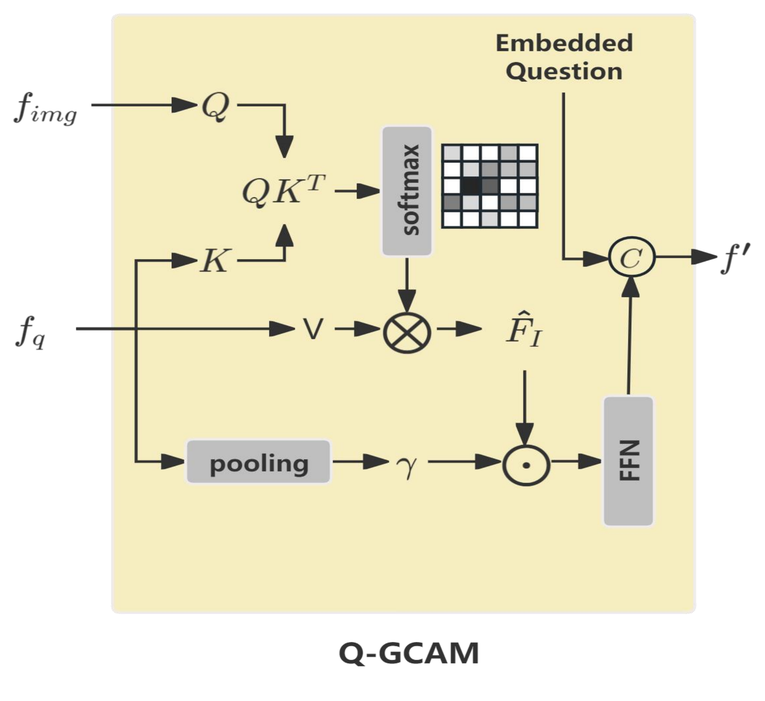
Durham University

30th ANNIVERSARY EMNLP 2025 Suzhou, China November 4-9 | 中国苏州 11月4日–9日







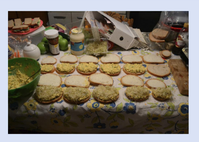
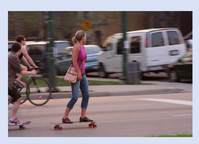


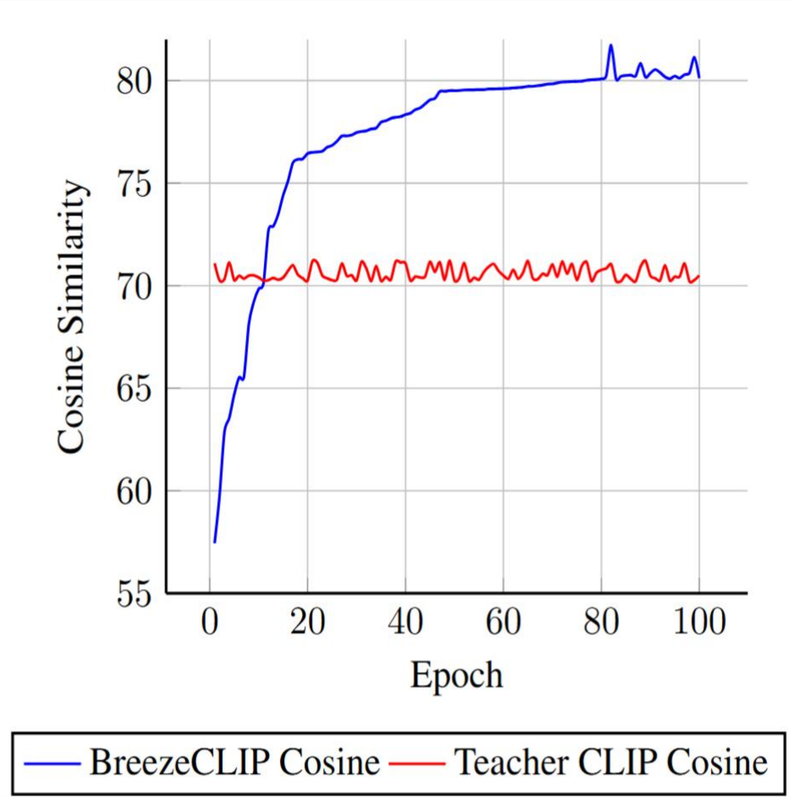
Contributions

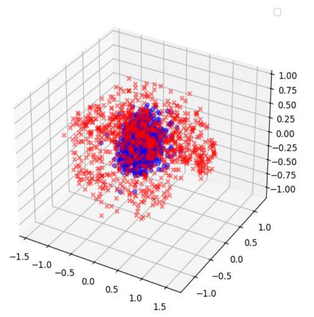
- We propose BcQLM, involving a compact BreezeCLIP by distillation learning and a Q-Gated Cross-Modal Fusion Module. Comprehensive experiments demonstrate that our BreezeCLIP is able to effectively preserve vision-language alignment capabilities under a tiny model setting.
- We propose a Q-Gated Cross-Modal Fusion Module to enable fine-grained and adaptive multimodal fusion.
- Our BreezeCLIP only contains 1.2B parameters (10\% of SoTA method), which achieves performance comparable to several standard-sized MLLMs with much higher parameter counts on VQA tasks.

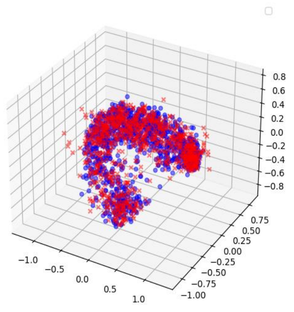
Results

Method	LLM	Param.	Res.	GQA	VQA^{V2}	VisWiz
BLIP-2 (2023)	Vicuna-13B	13.5	224	44.0	65.0	19.6
InstructBLIP (2023)	Vicuna-7B	7.5	224	49.2	–	34.5
InstructBLIP (2023)	Vicuna-13B	13.5	224	49.5	–	33.4
IDEFICS-9B (2023)	LLaMA-7B	9.0	224	38.4	50.9	35.5
IDEFICS-80B (2023)	LLaMA-65B	80.0	224	45.2	60.0	36.0
BcQLM (ours)	*Llama-3.2-1B*	**1.2**	*224*	**60.8**	**71.0**	**49.5**
Qwen-VL (2023)	Qwen-7B	9.6	448	59.3	78.8	35.2
Qwen-VL-Chat (2023)	Qwen-7B	9.6	448	57.5	78.2	38.9
LLaVA-1.5 (2024)	Vicuna-1.5-7B	7.3	336	62.0	78.5	50.0
LLaVA-1.5 (2024)	Vicuna-1.5-13B	13.3	336	**63.3**	80.0	53.6
VILA-7B (2024)	Llama-2-7B	7.0	336	62.3	79.9	57.8
VILA-13B (2024)	Llama-2-13B	13.0	336	**63.3**	**80.8**	**60.6**
BcQLM (ours)	*Llama-3.2-1B*	**1.2**	*336*	62.4	*78.7*	*56.1*

Input Image					
Input Question	Are there any forks in the bowl?	Of which color is the fur?	Is the van to the right of a bus?	Does the ground look bare?	Is there a bus or a window in the scene?
BcQLM(5%)	No, there is no fork in the bowl.	The fur is gray.	No, the van is to the right of a man.	No, the ground is snowy.	Yes, there is a window.
BcQLM(50%)	No, there is some batter in the bowl.	The fur is gray.	No, the van is to the right of people.	No, the ground is snowy.	Yes, there is a window.
BcQLM(100%)	No, there is a spoon in the bowl.	The fur is gray.	No, the van is to the right of the person.	No, the ground is snowy.	Yes, there are windows.
Ground Truth	No, there is a spoon in the bowl.	The fur is white.	No, the van is to the right of the person.	No, the ground is rocky.	Yes, there are windows.







More information

Github Paper Homepage