Pretrained Image-Text Models are Secretly Video Captioners

{Chunhui Zhang*[1], Yiren Jian*[2]}, Zhongyu Ouyang[1],
Soroush Vosoughi[1]

[1] Dartmouth College, [2] OpenAI

Reinforcement learning helped us achieve

a Top-2 ranking on the PaperWithCode leaderboard.

Recap

- Instruction tuning allowed the MLLMs to follow human instructions and improve performance on zero-shot tasks

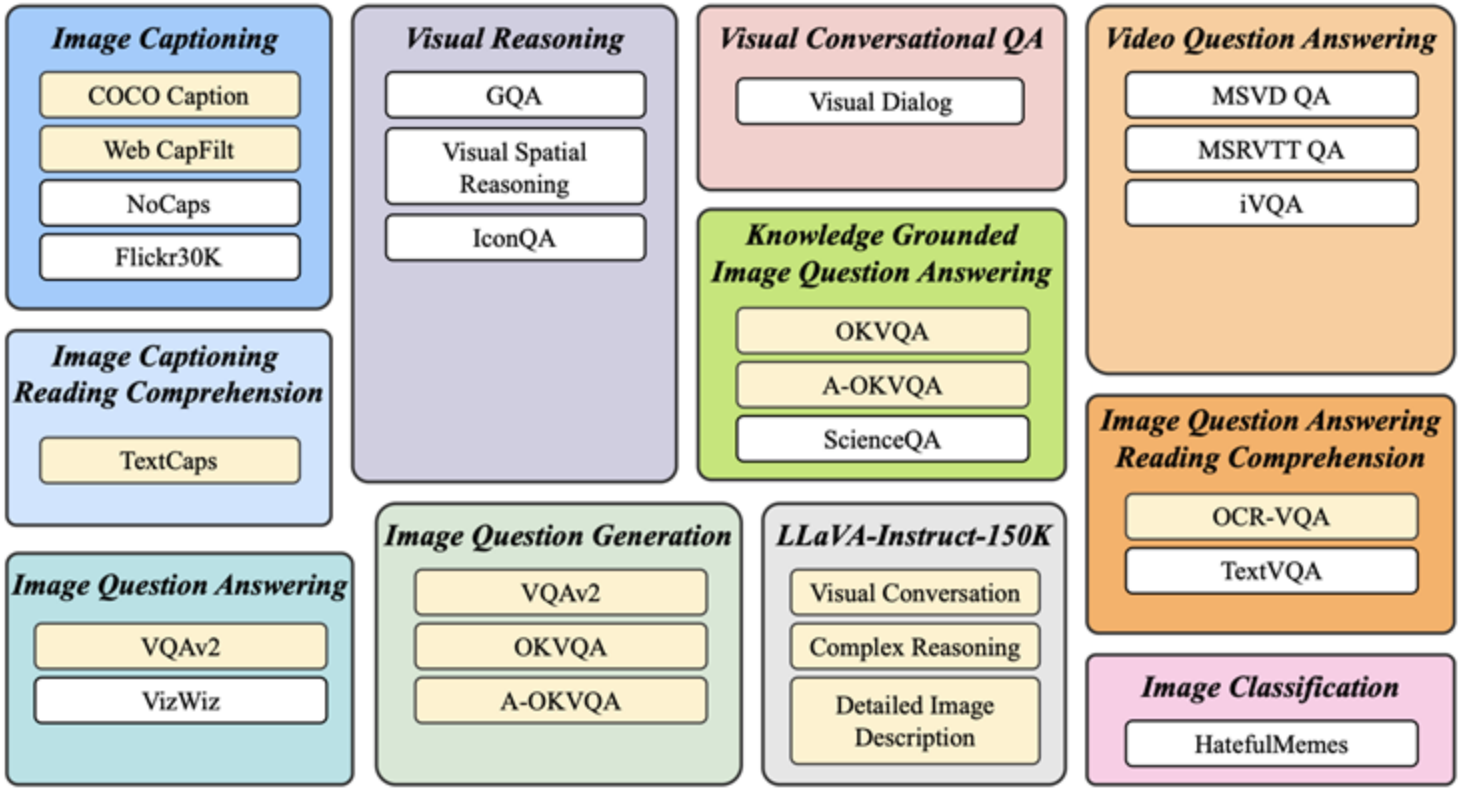


Figure copied from InstructBLIP paper

Challenges

- InstructBLIP excels at zero shot learning in image QA datasets but MSRVTT Video QA datasets.

	NoCaps	Flickr 30K	GQA	VSR	IconQA	TextVQA	Visdial	HM	VizWiz	SciQA image	MSVD QA	MSRVTT QA	iVQA
Flamingo-3B [4]	-	60.6	-	-	-	30.1	-	53.7	28.9	-	27.5	11.0	32.7
Flamingo-9B [4]	-	61.5	-	-	-	31.8	-	57.0	28.8	-	30.2	13.7	35.2
Flamingo-80B [4]	-	67.2	-	-	-	35.0	-	46.4	31.6	-	35.6	17.4	40.7
BLIP-2 (FlanT5$_{XL}$) [20]	104.5	76.1	44.0	60.5	45.5	43.1	45.7	53.0	29.8	54.9	33.7	16.2	40.4
BLIP-2 (FlanT5$_{XXL}$) [20]	98.4	73.7	44.6	68.2	45.4	44.1	46.9	52.0	29.4	64.5	34.4	17.4	45.8
BLIP-2 (Vicuna-7B)	107.5	74.9	38.6	50.0	39.7	40.1	44.9	50.6	25.3	53.8	18.3	9.2	27.5
BLIP-2 (Vicuna-13B)	103.9	71.6	41.0	50.9	40.6	42.5	45.1	53.7	19.6	61.0	20.3	10.3	23.5
InstructBLIP (FlanT5$_{XL}$)	119.9	**84.5**	48.4	64.8	50.0	46.6	46.6	56.6	32.7	70.4	43.4	25.0	53.1
InstructBLIP (FlanT5$_{XXL}$)	120.0	83.5	47.9	**65.6**	**51.2**	46.6	**48.5**	54.1	30.9	**70.6**	**44.3**	**25.6**	**53.8**
InstructBLIP (Vicuna-7B)	**123.1**	82.4	49.2	54.3	43.1	50.1	45.2	**59.6**	**34.5**	60.5	41.8	22.1	52.2
InstructBLIP (Vicuna-13B)	121.9	82.8	**49.5**	52.1	44.8	**50.7**	45.4	57.5	33.4	63.1	41.2	24.8	51.0

Motivation: Hypothesis

What if they jump over the doorway?

If the person were to jump over the doorway, they would be leaving the confines of their current location and venturing into the vastness of space. This could be a metaphorical representation of stepping out of one's comfort zone or embarking on a new adventure. However, it is important to note that such an action would be extremely dangerous and potentially life-threatening due to the harsh conditions of space.

Less focus on visual input

"Person might be caught in mid air on a sky"

*InstructBLIP may be using **shortcuts** to answer questions based on **correlations between questions and visual input***

Video captioning requires ***direct understanding***

InstructBLIP might be ignoring the temporal visual dynamics in video caption tasks

Question

Under resource constraints (model, data, supervision), how can image captioning models be efficiently repurposed for video captioning?

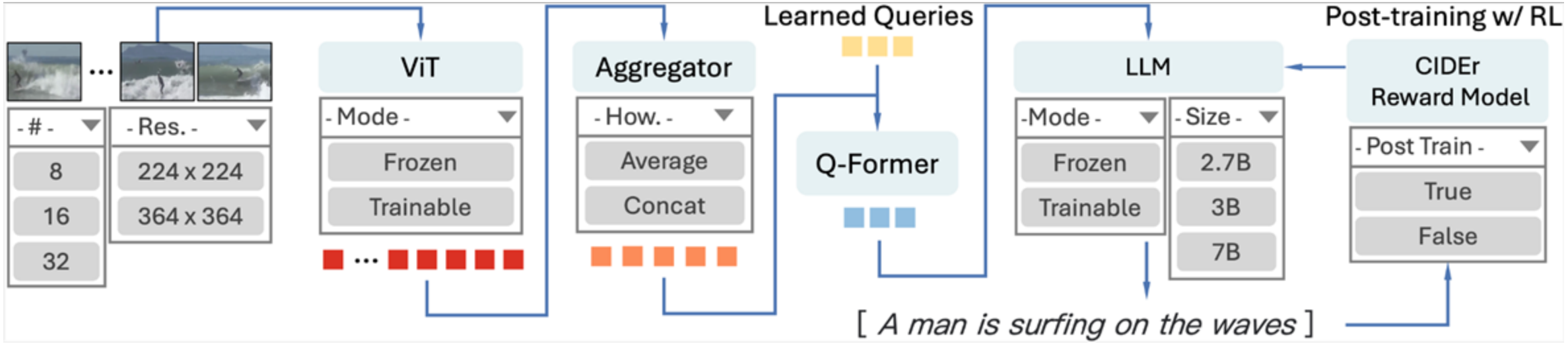
Hypothesis Test on Video

We first evaluated InstructBLIP on MSR-VTT-Caption using zero shot tasks for video captioning

Model	MSR-Video to Text [1]				Code	# video -text
	C.	M.	R.	B4.		-
IcoCap	60.2	31.1	64.9	47.0	No	-
MaMMUT	73.6	-	-	-	No	-
VideoCoCa	73.2	-	68.0	53.8	No	144.7M
VALOR	74.0	32.9	68.0	54.4	Yes	1.18M
VLAB	74.9	33.4	68.3	54.6	No	10.7M
GIT2	75.9	33.1	68.2	54.8	Yes	-
VAST	78.0	-	-	56.7	Yes	27M
mPLUG-2	80.0	34.9	70.1	57.8	Yes	2.5M
InstructBLIP	50.8	26.1	55.1	31.1	Yes	-

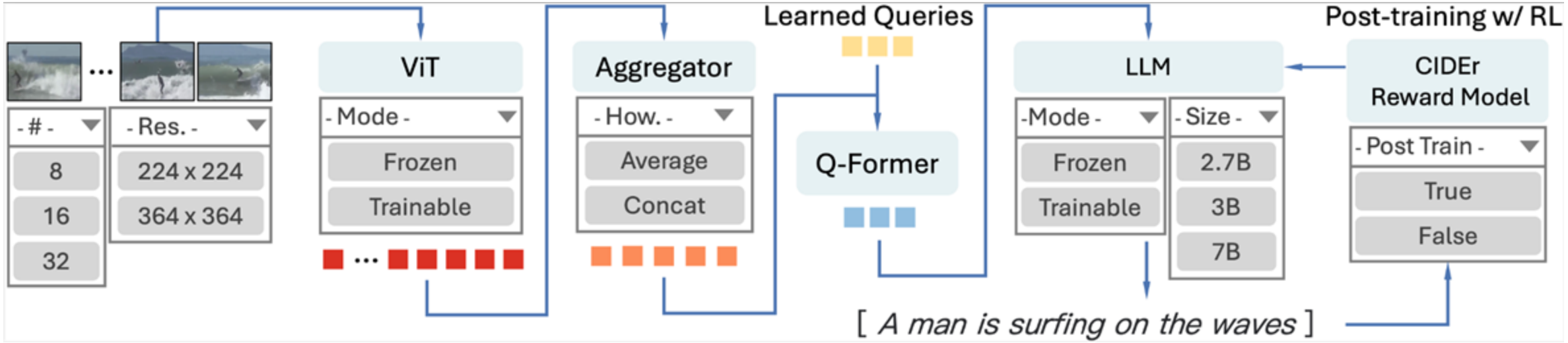
RL Post-training Pipeline

- We use CIDEr score as metric-based reward to supervise video captioning.



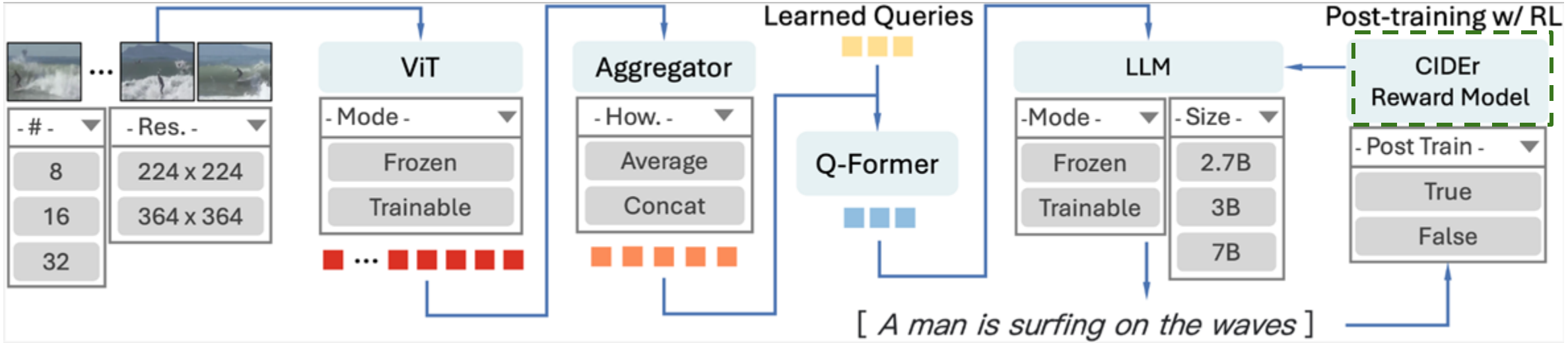
Model Architecture

- ViT, Q-Former as the modal connector, Flan-T5-XL as the language module
- Video inputs: 16 frames/video, 224x224 resolution.
- Query embeddings were concatenated before being fed into the language module



Optimal Configuration Summary

Fine-tune InstructBLIP model on MSR-VTT-Caption dataset, while 1)freezing ViT 2)Q-Former Only 3) Post-Training with Reinforcement Learning



Data Setup

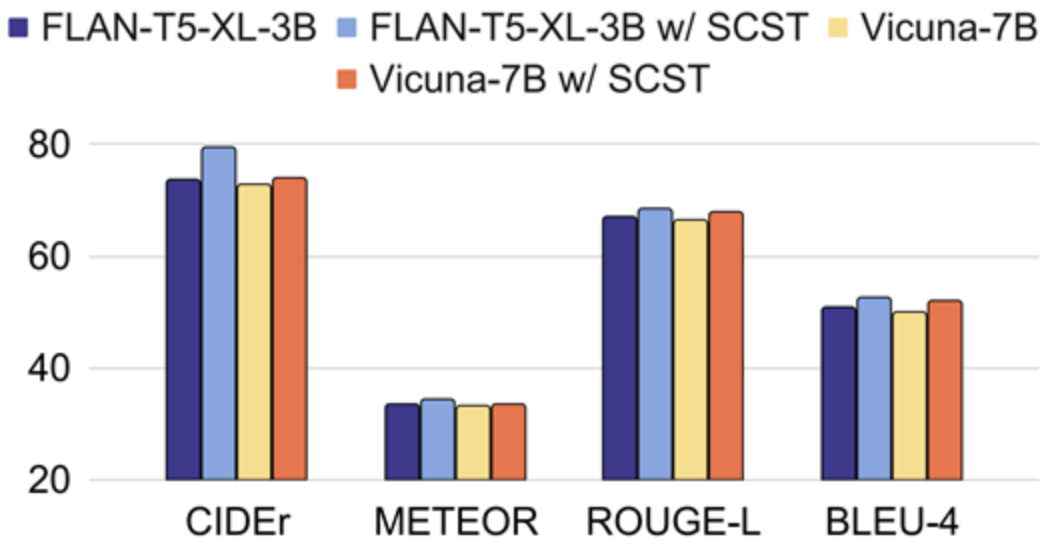
- Dataset Preparation
  - Utilized MSR-VTT-Caption dataset
    - 10,000 video clips across 20 categories
- Split data
  - 6,513 clips for training
  - 497 clips for validation
  - 2,990 clips for testing

We achieved **2nd best** against SoTA video captioners

Model	MSR-VTT-Caption [1]				Code	# video -text
	CIDEr	METEOR	ROUGE-L	BLEU-4		-
IcoCap [2]	60.2	31.1	64.9	47.0	No	-
MaMMUT [3]	73.6	-	-	-	No	-
VideoCoCa [4]	73.2	-	68.0	53.8	No	144.7M
VALOR [5]	74.0	32.9	68.0	54.4	Yes	1.18M
VLAB [6]	74.9	33.4	68.3	54.6	No	10.7M
GIT2 [7]	75.9	33.1	68.2	54.8	Yes	-
VAST [8]	78.0	-	-	56.7	Yes	27M
mPLUG-2 [9]	80.0	34.9	70.1	57.8	Yes	2.5M
InstructBLIP [10]	50.8	26.1	55.1	31.1	Yes	-
Ours	79.5	34.2	68.3	52.4	Yes	6K

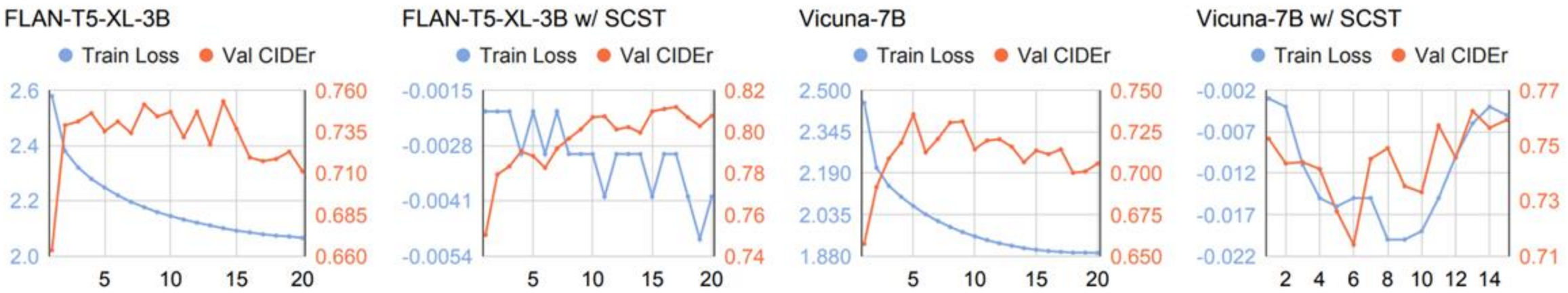
Key Findings

- Reinforcement Supervision:
  - Reinforcement learning (SCST) aligns captions with human preferences
  - Improved CIDEr scores by 3.4-6.5%



Key Findings

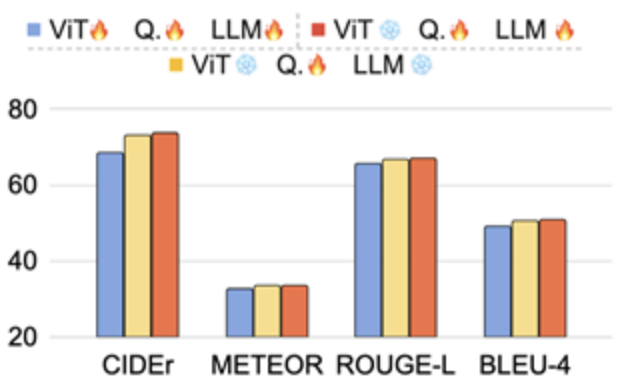
- Reinforcement Supervision:
  - Reinforcement learning (SCST) aligns captions with human preferences
  - Improved CIDEr scores by 3.4-6.5%

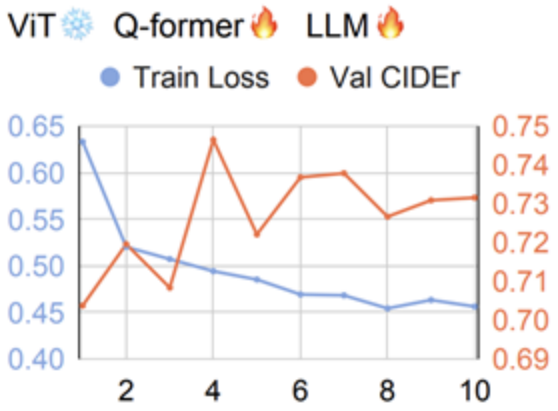


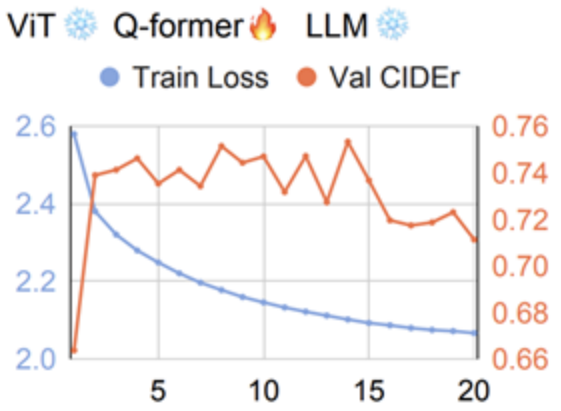
Key Findings

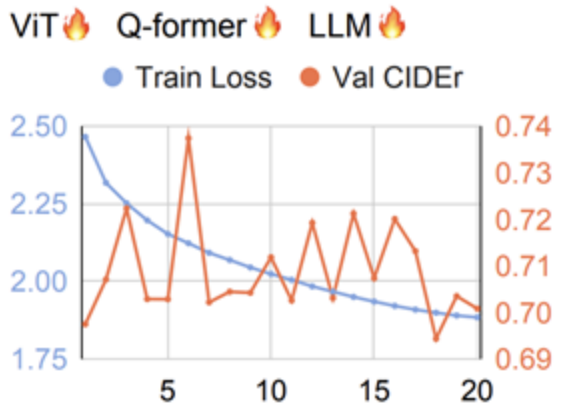
- Model Scale:

  Trainability hierarchy: Q-Former > LLM > ViT





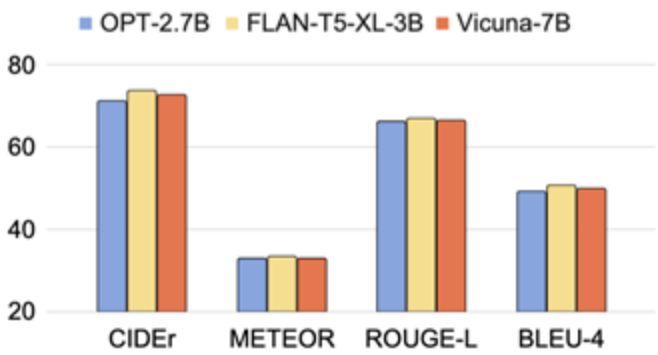


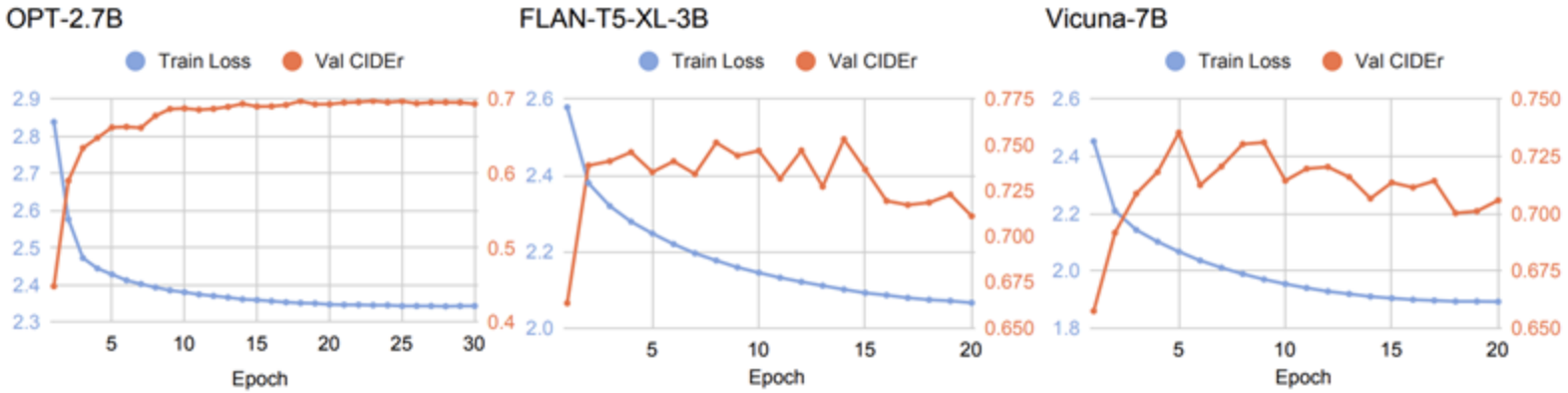


Key Findings

- Model Scale:

  Mid-sized LLMs (e.g. Flan-T5-XL-3B) work best for video captioning

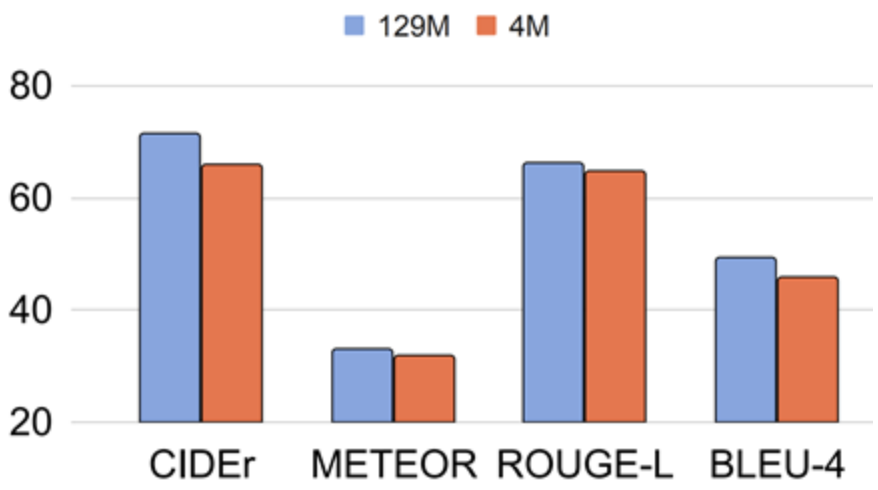




Key Findings

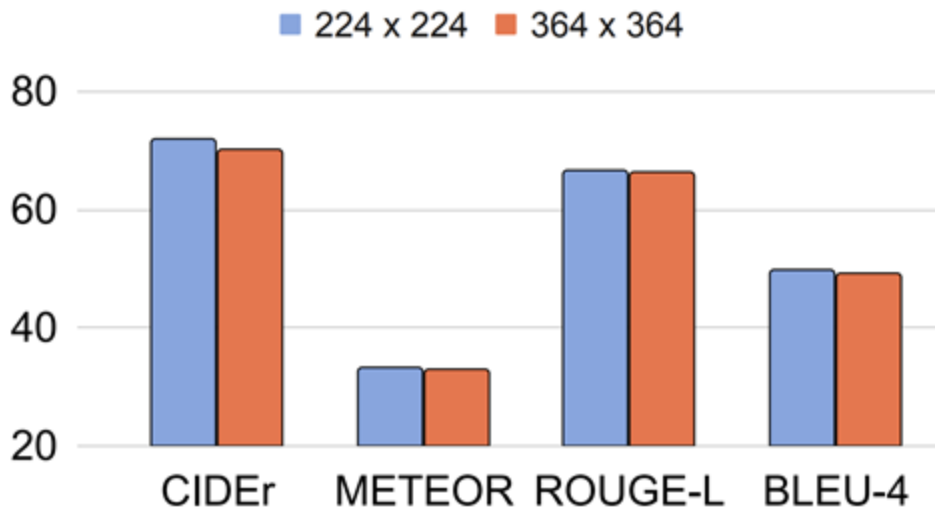
- Data Efficiency:

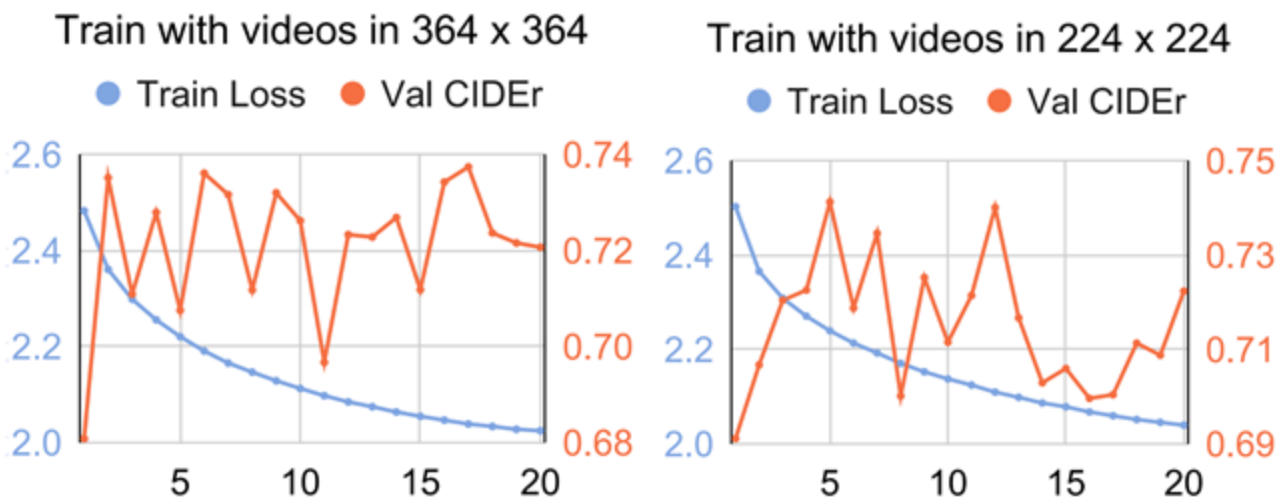
  Larger image-text pretraining datasets improve performance



Key Findings

- Data Efficiency:
  Lower resolution (224x224) works efficiently

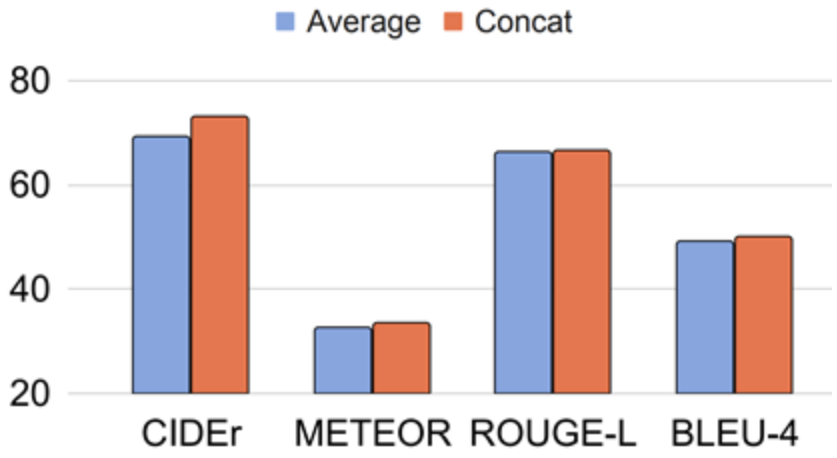


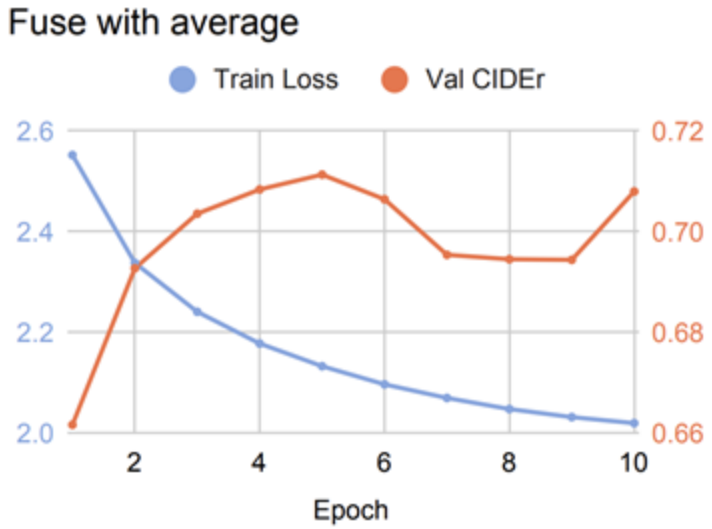


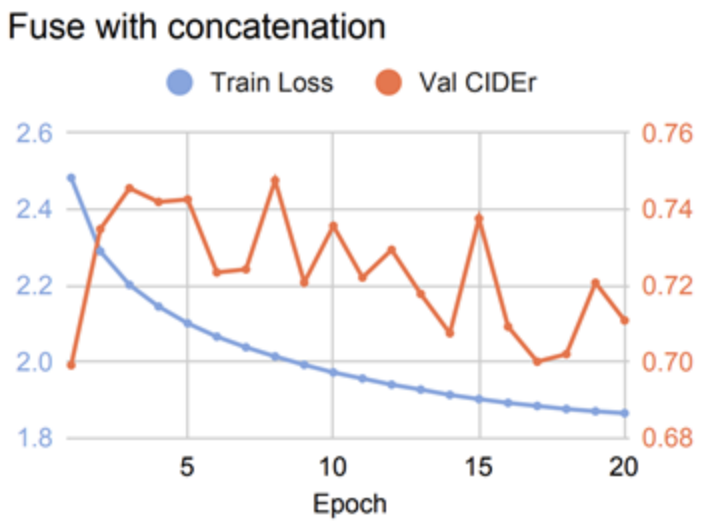
Key Findings

- Data Efficiency:

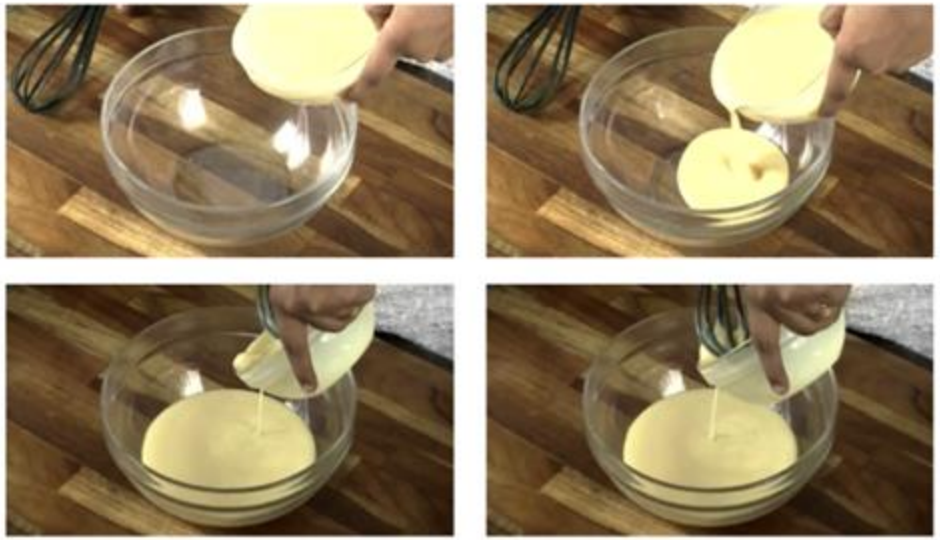
  Frame concatenation better captures temporality than averaging







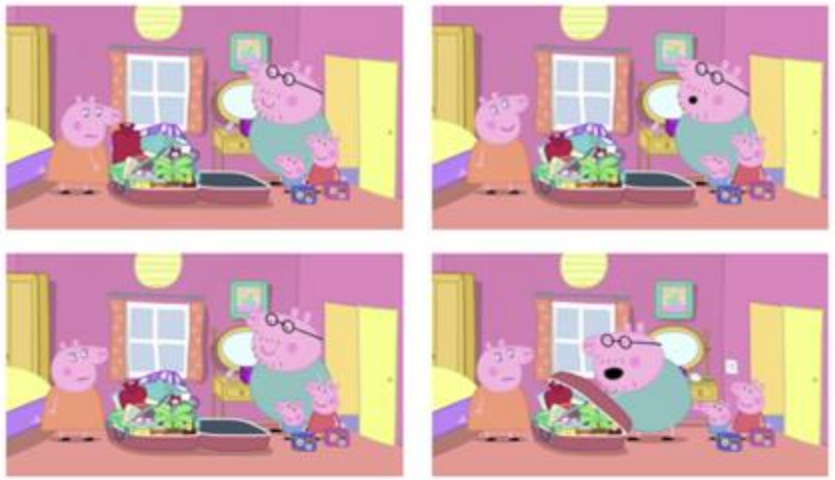
Qualitative Comparison: InstructBLIP v.s. Ours



((a)) **Baseline:** A hand pouring cream into a glass bowl. A hand whisking the cream in the bowl. The cream has been whipped to a fluffy consistency. A hand mixing the cream with a whisk. **Ours:** A woman introduces and showcases her ingredients, including readily available condensed milk, as she places it into a large bowl.

((b)) **Baseline:** Person driving a car. Person holding a smartphone while driving. Person's hands on the steering wheel. Person's face is blurred while driving. **Ours:** A man in sunglasses drives a car with the roar of a high-speed engine in the background, expressing his desire for unlimited fuel to keep driving into the sunset forever.

Qualitative Comparison: InstructBLIP v.s. Ours



((c)) **Baseline:** A pig wearing glasses is holding a tray with various objects on it, and two smaller pig characters are gathered around the tray. The pig is standing in a pink room with a window and a door, and there are some items scattered around the room. The pig is interacting with the two smaller pig characters in a playful and engaging manner. **Ours:** A group of cartoon characters, including piggy ones, prepares for a trip. One character double-checks if they have packed everything, while another emphasizes the importance of each item.

((d)) **Baseline:** Getting ready to take flight. Mid-air magic. Landing with style. Cruising in comfort. **Ours:** A young boy skateboards at a skate park, explaining skateboarding techniques and demonstrating how to perform a trick by using your hands to grab the nose of the board for better control.

Summary

- Identified **shortcut** in current InstructBLIP's visual understanding
- Key factors for recycling instructBLIP to video captioning
- Achieved **top** performance with **minimal** resources by **RL**