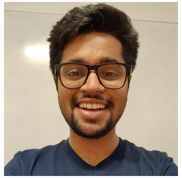
Contrast and Classify: Alternate Training for Robust VQA


Yash Kant


Abhinav Moudgil


Dhruv Batra

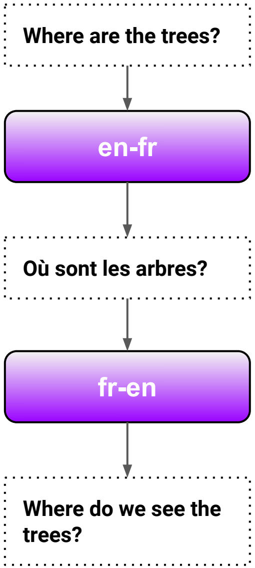

Devi Parikh


Harsh Agrawal

Inconsistency in VQA Models

	Prediction
What is in the basket?	banana
What is contained in the basket?	pizza
What can be seen inside the basket?	remote
What does the basket mainly contain?	paper

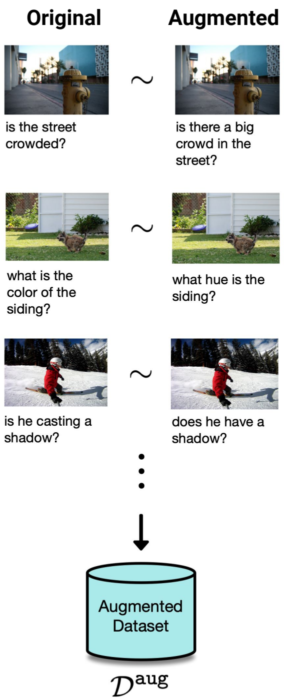
Data Augmentation with Back Translation



- Related works have proposed to augment the training data with paraphrases.
- One way to generate paraphrases is to use Back Translation using a pair pre-trained MT models[1].
- Back Translation involves converting converting the original question to a second language and translating it back to english using two pretrained MT models.

1. Marian MT: https://huggingface.co/transformers/model_doc/marian.html

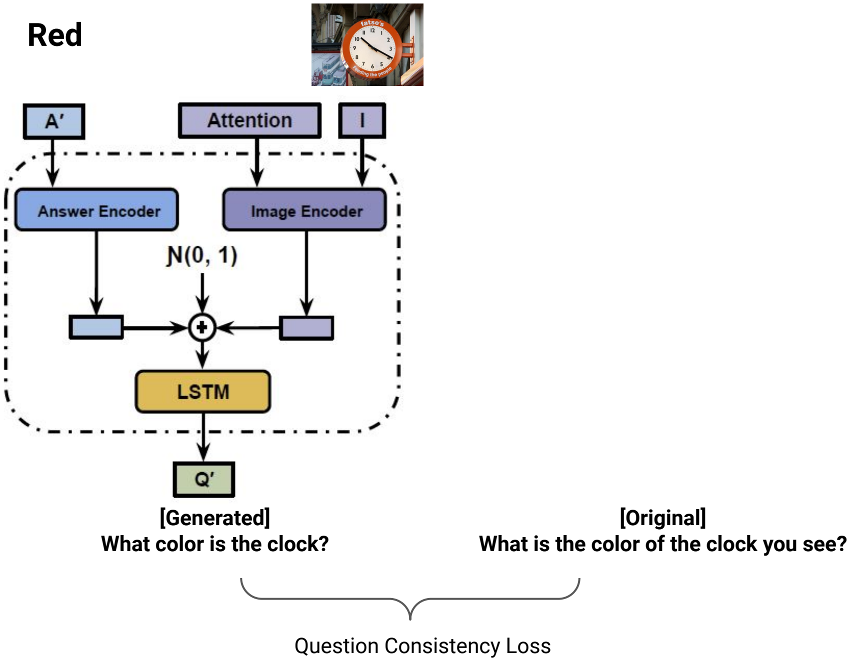
Data Augmentation with Back Translation



- Related works have proposed to augment the training data with paraphrases.
- One way to generate paraphrases is to use Back Translation using a pair pre-trained MT models[1].
- Back Translation involves converting converting the original question to a second language and translating it back to english using two pretrained MT models.
- We augment the VQA dataset with rephrased samples.

1. Marian MT: https://huggingface.co/transformers/model_doc/marian.html

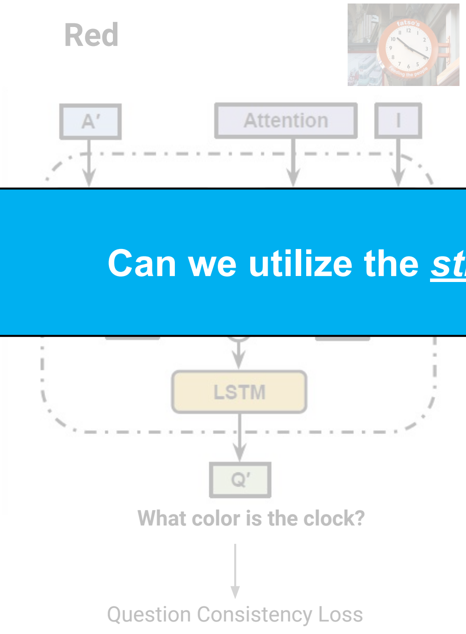
Data Augmentation with VQG



VQG Module

- Another way to generate paraphrases is by using a visual question generation (VQG) module.
- Given an image-answer pair from the VQA dataset it attempts to generate a rephrasing of the original question.
- To ensure that the generated question is a paraphrase of the original it is trained using a Question Consistency loss.

Data Augmentation with VQG



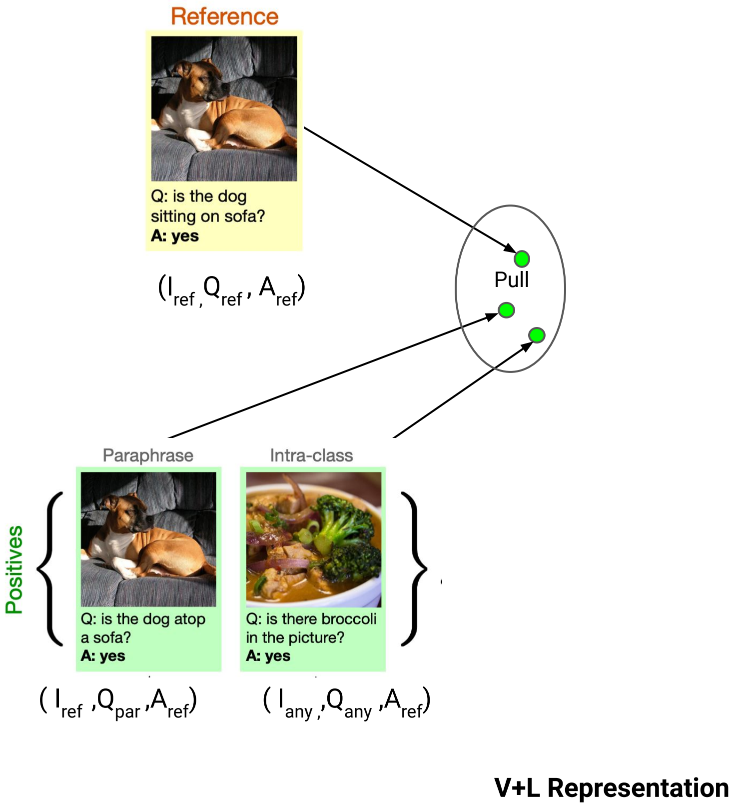
VQG Module

- Another way to generate paraphrases is by using a VQG module.
- Given an image-answer pair it generates a question.

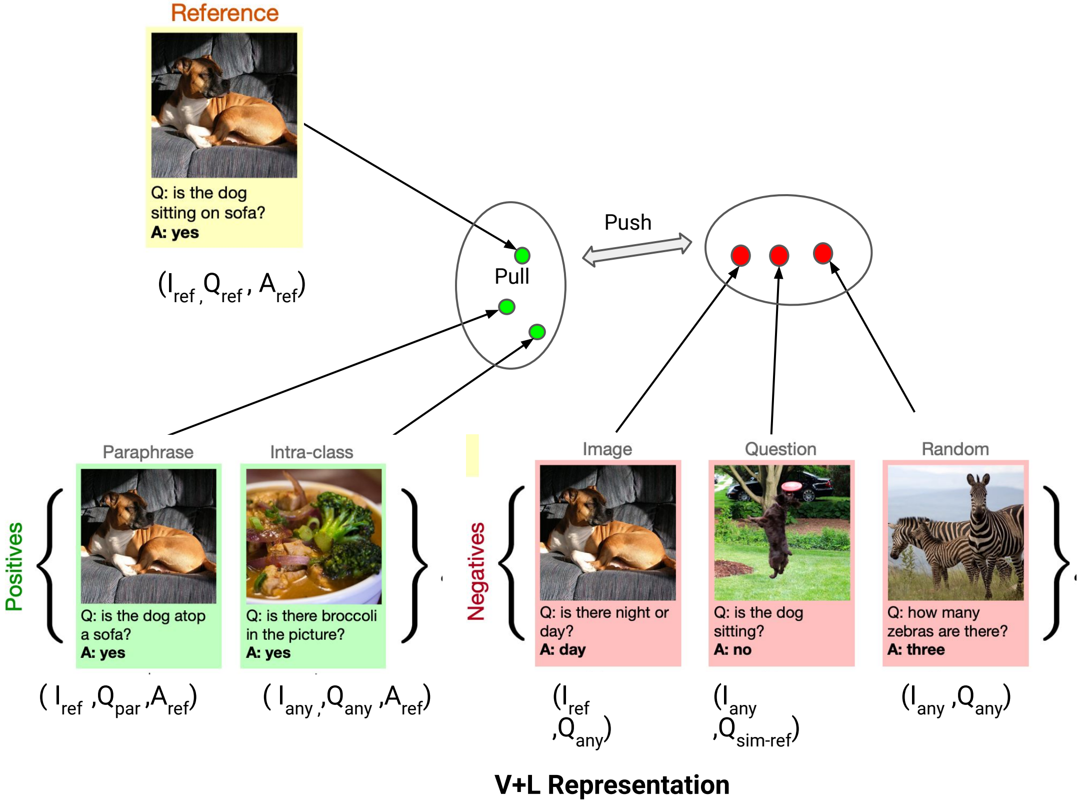
Can we utilize the *structure* in our augmented data to learn better?

Supervised Contrastive Loss

- SCL helps us to pull closer the joint V+L representations of reference and paraphrased sample closer.
- **Positives:** In addition to paraphrased samples, we also pull closer samples with same ground truth.

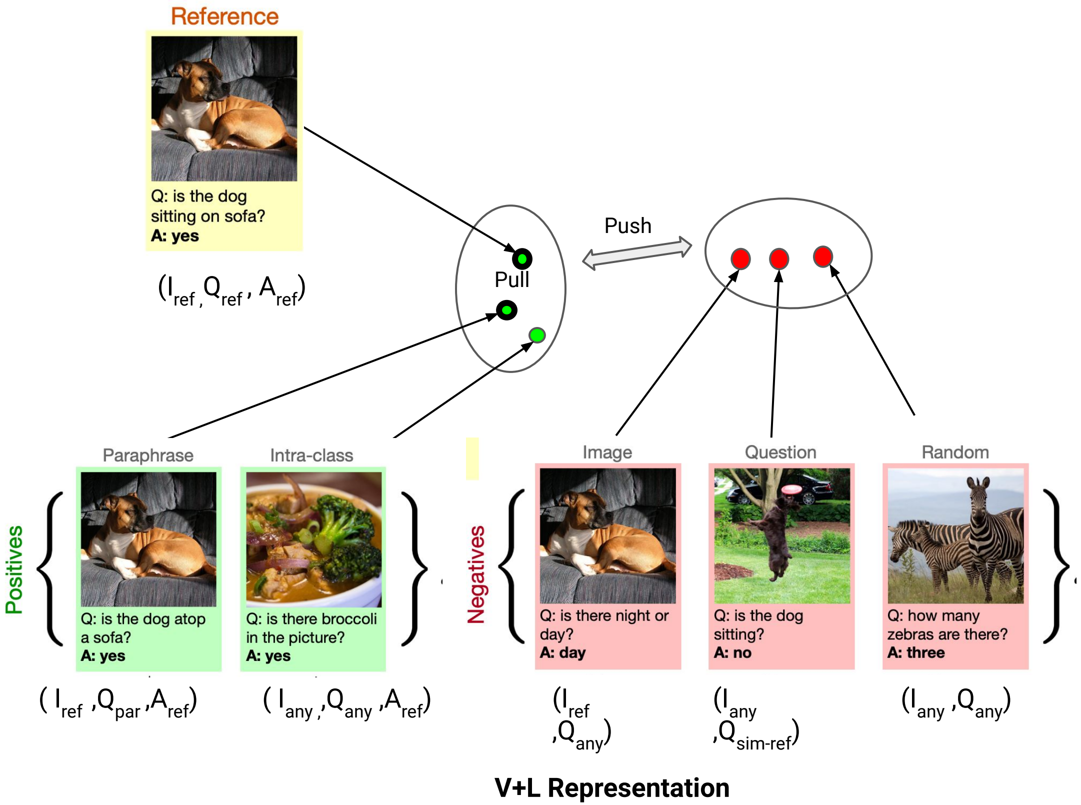


Supervised Contrastive Loss



- SCL helps us to pull closer the joint V+L representations of reference and paraphrased sample closer.
- **Positives:** In addition to paraphrased samples, we also pull closer samples with same ground truth.
- While pulling positives closer we also push negatives apart.
- **Negatives:** We categorize the negatives in three types — Image, Question and Random

Scaled Supervised Contrastive Loss



- VQA is a skewed dataset, and while creating batches for SCL the number of intra-class positives outnumber the paraphrased positives by a big margin for common answer categories ('yes', 'no', etc.).
- **Scaling:** To mitigate the above problem we scale the loss between the reference and the corresponding paraphrased positive sample.
- SCL batches are curated by sampling from sets of positive and negative types (not random).

Training Scheme

SSCL Iteration:

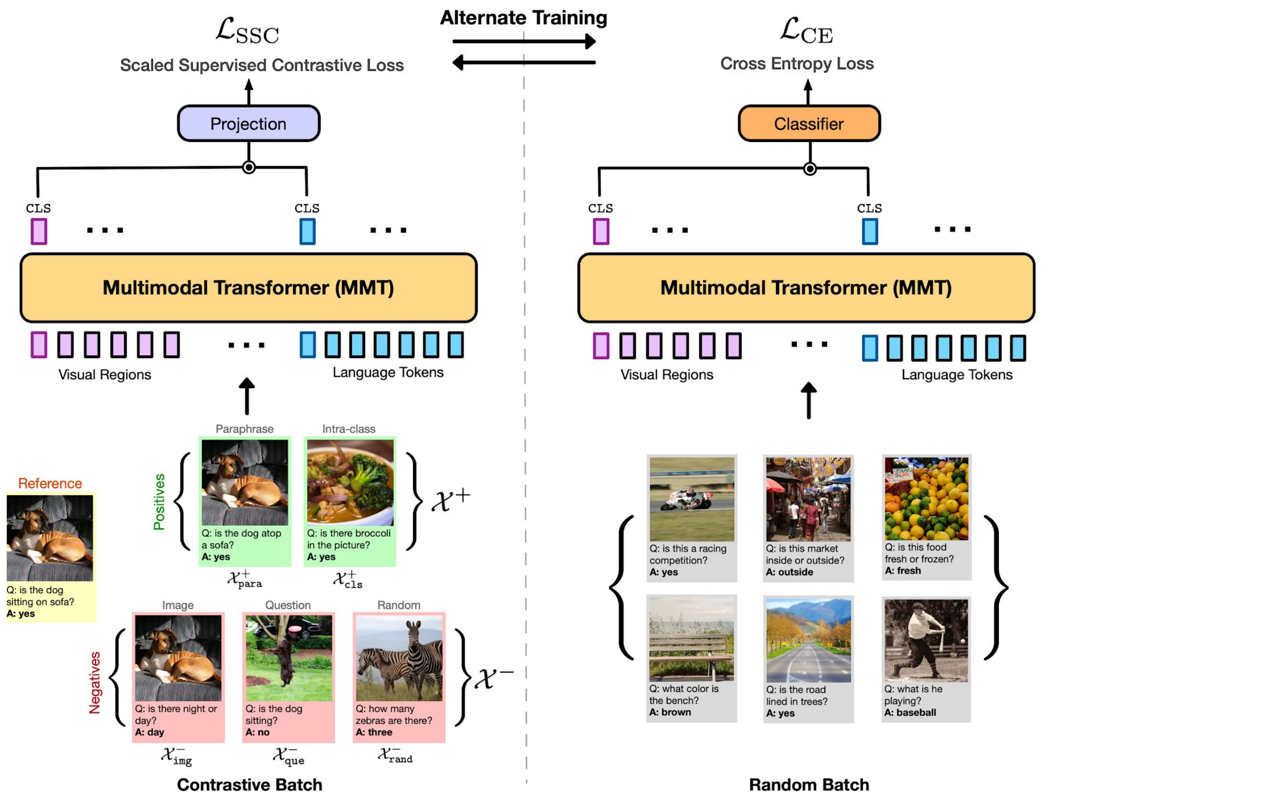
1. Curate batch with positives and negatives
2. Minimize the Scaled Supervised Contrastive Loss

CE Iteration:

1. Randomly sample batch
2. Minimize the Cross Entropy Loss

- **Joint:** We use a linear combination of the SCL and Cross Entropy losses as our objective. *Drawback of this approach is that either Cross Entropy is forced to operate on batches curated for SCL or vice-versa.*
- **Alternate:** We alternately optimize for SCL and Cross Entropy across different iterations. This allows using curated batches for SCL and random batches for Cross Entropy.

Contrast and Classify: Alternate Training for Robust VQA (ConCAT)



Experimental Results

Previous Work

	Model	DA	Scaling	N-Type	Consensus Score(k) k=1	k=2	k=3	k=4	VQA Score val	test-dev	test-std
1	Pythia (2018)	-	-	-	63.43	52.03	45.94	39.49	65.78	68.43	-
2	BAN (2018)	-	-	-	64.88	53.08	47.45	39.87	66.04	69.64	-
3	Pythia + CC (2019)	-	-	-	64.36	55.45	50.92	44.30	66.03	68.88	-
4	BAN + CC (2019)	-	-	-	65.77	56.94	51.76	48.18	66.77	**69.87**	-
5	MMT + CE	-	-	-	67.74	59.82	55.10	51.82	66.46	-	-
6	MMT + CE	VQG	-	-	66.53	59.26	54.92	51.85	64.50	-	-
7	MMT + ConCAT	VQG	✓	RQI	66.49	59.55	55.33	52.31	64.74	-	-
8	MMT + CE	BT	-	-	67.58	60.04	55.53	52.36	66.31	69.51	69.22
9	MMT + (SCL→CE)	BT	✗	R	65.34	57.39	52.63	49.20	64.21	-	-
10	MMT + (CE + SCL)	BT	✗	R	66.95	59.70	55.32	52.20	65.10	-	-
11	MMT + ConCAT	BT	✗	R	68.35	60.97	56.49	53.30	66.73	-	-
12	MMT + ConCAT	BT	✓	R	68.19	60.92	56.53	53.42	66.62	-	-
13	MMT + ConCAT	BT	✓	RQ	68.41	61.24	56.88	53.77	66.97	-	-
14	MMT + ConCAT	BT	✓	RI	68.47	61.28	56.91	53.79	66.93	-	-
15	MMT + ConCAT	BT	✗	RQI	68.20	60.90	56.49	53.36	66.60	-	-
16	MMT + ConCAT	BT	✓	RQI	**68.62**	**61.42**	**57.08**	**53.99**	**66.98**	69.80	**70.00**

Experimental Results

		Model	DA	Scaling	N-Type	Consensus Score(k)				VQA Score		
						k=1	k=2	k=3	k=4	val	test-dev	test-std
Previous Work	1	Pythia (2018)	-	-	-	63.43	52.03	45.94	39.49	65.78	68.43	-
	2	BAN (2018)	-	-	-	64.88	53.08	47.45	39.87	66.04	69.64	-
	3	Pythia + CC (2019)	-	-	-	64.36	55.45	50.92	44.30	66.03	68.88	-
	4	BAN + CC (2019)	-	-	-	65.77	56.94	51.76	48.18	66.77	**69.87**	-
Arch. Change	5	MMT + CE	-	-	-	67.74	59.82	55.10	51.82	66.46	-	-
	6	MMT + CE	VQG	-	-	66.53	59.26	54.92	51.85	64.50	-	-
	7	MMT + ConCAT	VQG	✓	RQI	66.49	59.55	55.33	52.31	64.74	-	-
	8	MMT + CE	BT	-	-	67.58	60.04	55.53	52.36	66.31	69.51	69.22
	9	MMT + (SCL→CE)	BT	✗	R	65.34	57.39	52.63	49.20	64.21	-	-
	10	MMT + (CE + SCL)	BT	✗	R	66.95	59.70	55.32	52.20	65.10	-	-
	11	MMT + ConCAT	BT	✗	R	68.35	60.97	56.49	53.30	66.73	-	-
	12	MMT + ConCAT	BT	✓	R	68.19	60.92	56.53	53.42	66.62	-	-
	13	MMT + ConCAT	BT	✓	RQ	68.41	61.24	56.88	53.77	66.97	-	-
	14	MMT + ConCAT	BT	✓	RI	68.47	61.28	56.91	53.79	66.93	-	-
	15	MMT + ConCAT	BT	✗	RQI	68.20	60.90	56.49	53.36	66.60	-	-
	16	MMT + ConCAT	BT	✓	RQI	**68.62**	**61.42**	**57.08**	**53.99**	**66.98**	69.80	**70.00**

Experimental Results

		Model	DA	Scaling	N-Type	Consensus Score(k)					VQA Score	
						k=1	k=2	k=3	k=4	val	test-dev	test-std
Previous Work	1	Pythia (2018)	-	-	-	63.43	52.03	45.94	39.49	65.78	68.43	-
	2	BAN (2018)	-	-	-	64.88	53.08	47.45	39.87	66.04	69.64	-
	3	Pythia + CC (2019)	-	-	-	64.36	55.45	50.92	44.30	66.03	68.88	-
	4	BAN + CC (2019)	-	-	-	65.77	56.94	51.76	48.18	66.77	**69.87**	-
Arch. Change	5	MMT + CE	-	-	-	67.74	59.82	55.10	51.82	66.46	-	-
	6	MMT + CE	VQG	-	-	66.53	59.26	54.92	51.85	64.50	-	-
	7	MMT + ConCAT	VQG	✓	RQI	66.49	59.55	55.33	52.31	64.74	-	-
Baseline	8	MMT + CE	BT	-	-	67.58	60.04	55.53	52.36	66.31	69.51	69.22
	9	MMT + (SCL→CE)	BT	✗	R	65.34	57.39	52.63	49.20	64.21	-	-
	10	MMT + (CE + SCL)	BT	✗	R	66.95	59.70	55.32	52.20	65.10	-	-
	11	MMT + ConCAT	BT	✗	R	68.35	60.97	56.49	53.30	66.73	-	-
	12	MMT + ConCAT	BT	✓	R	68.19	60.92	56.53	53.42	66.62	-	-
	13	MMT + ConCAT	BT	✓	RQ	68.41	61.24	56.88	53.77	66.97	-	-
	14	MMT + ConCAT	BT	✓	RI	68.47	61.28	56.91	53.79	66.93	-	-
	15	MMT + ConCAT	BT	✗	RQI	68.20	60.90	56.49	53.36	66.60	-	-
	16	MMT + ConCAT	BT	✓	RQI	**68.62**	**61.42**	**57.08**	**53.99**	**66.98**	69.80	**70.00**

Experimental Results

		Model	DA	Scaling	N-Type	Consensus Score(k)				VQA Score		
						k=1	k=2	k=3	k=4	val	test-dev	test-std
Previous Work	1	Pythia (2018)	-	-	-	63.43	52.03	45.94	39.49	65.78	68.43	-
	2	BAN (2018)	-	-	-	64.88	53.08	47.45	39.87	66.04	69.64	-
	3	Pythia + CC (2019)	-	-	-	64.36	55.45	50.92	44.30	66.03	68.88	-
	4	BAN + CC (2019)	-	-	-	65.77	56.94	51.76	48.18	66.77	**69.87**	-
Arch. Change	5	MMT + CE	-	-	-	67.74	59.82	55.10	51.82	66.46	-	-
	6	MMT + CE	VQG	-	-	66.53	59.26	54.92	51.85	64.50	-	-
	7	MMT + ConCAT	VQG	✓	RQI	66.49	59.55	55.33	52.31	64.74	-	-
Baseline	8	MMT + CE	BT	-	-	67.58	60.04	55.53	52.36	66.31	69.51	69.22
Ablations	9	MMT + (SCL→CE)	BT	✗	R	65.34	57.39	52.63	49.20	64.21	-	-
	10	MMT + (CE + SCL)	BT	✗	R	66.95	59.70	55.32	52.20	65.10	-	-
	11	MMT + ConCAT	BT	✗	R	68.35	60.97	56.49	53.30	66.73	-	-
	12	MMT + ConCAT	BT	✓	R	68.19	60.92	56.53	53.42	66.62	-	-
	13	MMT + ConCAT	BT	✓	RQ	68.41	61.24	56.88	53.77	66.97	-	-
	14	MMT + ConCAT	BT	✓	RI	68.47	61.28	56.91	53.79	66.93	-	-
	15	MMT + ConCAT	BT	✗	RQI	68.20	60.90	56.49	53.36	66.60	-	-
	16	MMT + ConCAT	BT	✓	RQI	**68.62**	**61.42**	**57.08**	**53.99**	**66.98**	69.80	**70.00**

Thank You