

# Learning Distinct and Representative Modes for Image Captioning

Qi Chen, Chaorui Deng, Qi Wu

Australian Institute for Machine Learning (AIML), University of Adelaide

# Problem Clarification



## **Captions generated by humans:**

*A man holding a white frisbee while standing on a field.*

*A man standing outside holding a frisbee in his hands.*

*A man is outside holding onto a frisbee disc.*

*A smiling man in the park with a frisbee.*

*A man is having a good time playing frisbee.*

## **Captions generated by existing models:**

**Transformer:** *A man holding a frisbee while standing in a field.*

**AoANet:** *A man holding a frisbee in a field.*

## **Captions generated by DML (Ours):**

**Mode-7:** *There is a man playing with a frisbee.*

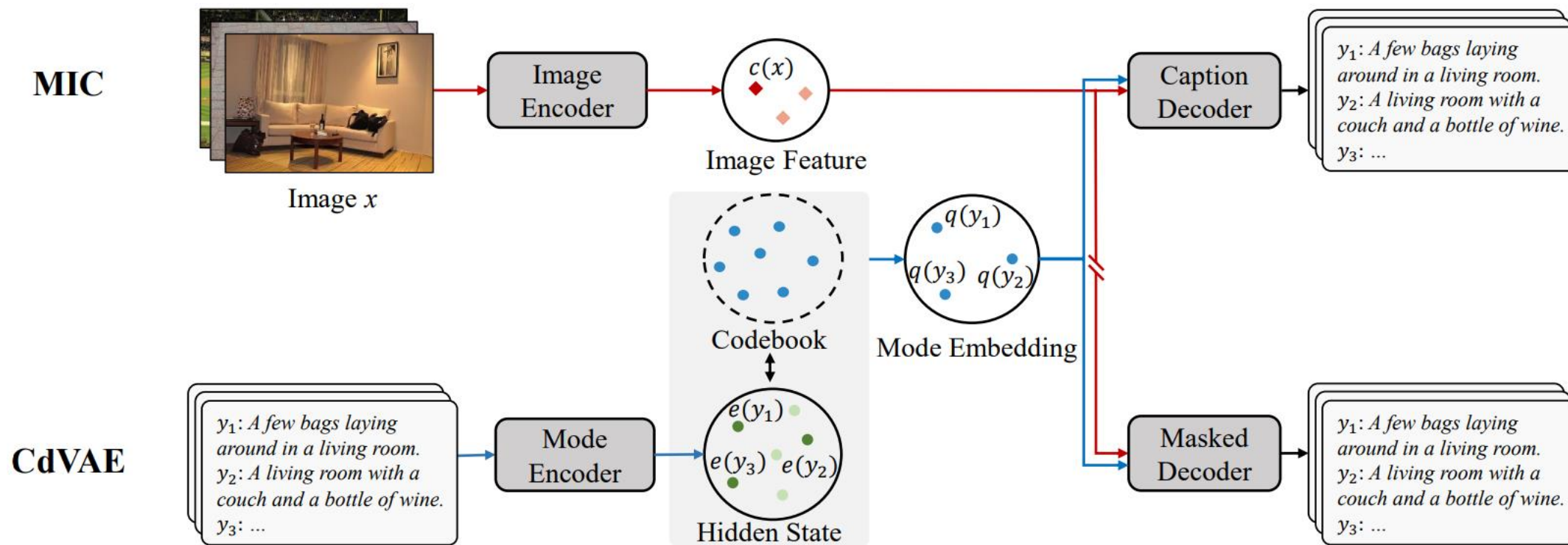
**Mode-32:** *Man in a blue shirt throwing a white frisbee.*

**Mode-43:** *A close up of a person with a frisbee.*

**Mode-3:** *A young man holding a white frisbee in his right hand.*

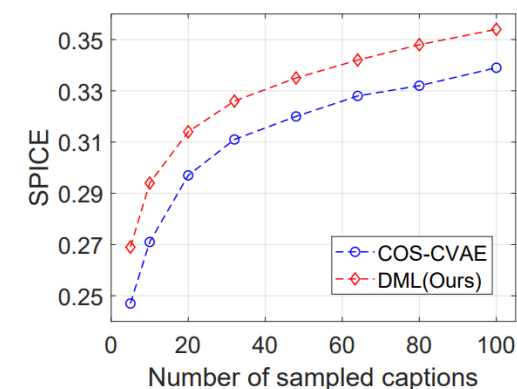
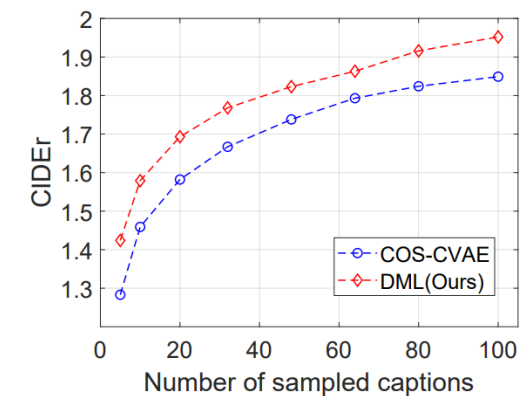
**Mode-58:** *A man is about to throw a frisbee.*

# Method: Discrete Mode Learning (DML)






# Experiment: Quality

Method	#Sample	B@1	B@2	B@3	B@4	R	M	C	S
Div-BS [45]	20	0.837	0.687	0.538	0.383	0.653	0.357	1.405	0.269
POS [14]		0.874	0.737	0.593	0.449	0.678	0.365	1.468	0.277
AG-CVAE [47]		0.834	0.698	0.573	0.471	0.638	0.309	1.259	0.244
Seq-CVAE [3]		0.870	0.727	0.591	0.445	0.671	0.356	1.448	0.279
COS-CVAE [31]		0.903	0.771	0.640	0.500	0.706	0.387	1.624	0.295
AoANet-DML (Ours)		<b>0.917</b>	<b>0.799</b>	<b>0.682</b>	<b>0.554</b>	<b>0.734</b>	<b>0.418</b>	<b>1.734</b>	<b>0.328</b>
Transformer-DML (Ours)		<u>0.915</u>	<u>0.788</u>	<u>0.663</u>	<u>0.526</u>	<u>0.726</u>	<u>0.417</u>	<u>1.704</u>	<u>0.325</u>
Div-BS [45]	100	0.846	0.698	0.555	0.402	0.666	0.372	1.448	0.290
POS [14]		0.909	0.787	0.672	0.550	0.725	0.409	1.661	0.311
AG-CVAE [47]		0.883	0.767	0.654	0.557	0.690	0.345	1.517	0.277
Seq-CVAE [3]		0.922	0.803	0.691	0.575	0.733	0.410	1.695	0.320
LNFM [30]		0.920	0.802	0.695	0.597	0.729	0.402	1.705	0.316
COS-CVAE [31]		0.942	0.842	0.739	0.633	0.770	0.450	1.893	0.339
AoANet-DML (Ours)		<b>0.947</b>	<b>0.850</b>	<b>0.752</b>	<b>0.652</b>	<b>0.782</b>	<b>0.479</b>	<b>1.960</b>	<b>0.356</b>
Transformer-DML (Ours)	<u>0.946</u>	<u>0.849</u>	<u>0.750</u>	<u>0.649</u>	<u>0.780</u>	<u>0.474</u>	<u>1.953</u>	<u>0.354</u>	



# Experiment: Diversity

Methods	LNFM [30]	COS-CVAE [31]	Seq-CVAE [3]	Transformer-BS	Transformer-DML
Div-1 (↑)	0.37	0.39	0.33	0.21	<b>0.43</b>
Div-2 (↑)	0.50	0.57	0.48	0.29	<b>0.59</b>
SelfCIDEr (↑)	-	0.79	-	0.57	<b>0.83</b>
mBLEU (↓)	0.64	<b>0.53</b>	0.64	0.78	0.54

Image			
<b>Mode 7</b>	<i>There is a man in the middle of a field playing with a frisbee</i>	<i>There is a man riding a motorcycle down the street</i>	<i>There is a man on a skateboard holding a can</i>
<b>Mode 32</b>	<i>Man in red shirt throwing a white frisbee</i>	<i>Man on a yellow motorcycle driving down the road</i>	<i>Man on skateboard with a beer on the ground</i>
<b>Mode 43</b>	<i>A close up of a person with a frisbee</i>	<i>A close up of a person riding a motorcycle on a road</i>	<i>A close up of a person on a skateboard</i>
<b>Mode 3</b>	<i>A young man holding a white frisbee on top of a green field</i>	<i>A man riding a yellow motorcycle with a yellow helmet on</i>	<i>A young man holding a bottle of water while standing on a skateboard</i>
<b>Mode 58</b>	<i>A man getting ready to throw a frisbee</i>	<i>A man that is sitting on a motorcycle</i>	<i>A person on a court with a skateboard</i>

Thank you for your attention!