

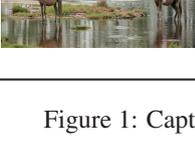
Image	Generated captions	Ground truths
	<p>Base model: A wooden bench sitting on top of a wooden bench.</p> <p>Encoder (w/ refining, w/o AoA): A box of books sitting on top of a wooden bench.</p> <p>Encoder (w/ refining, w/ AoA): A book sitting on top of a wooden bench.</p> <p>LSTM decoder (H=1): A wooden bench sitting next to a wooden bench.</p> <p>AoA decoder (H=1): A wooden bench sitting on top of a wooden bench.</p> <p>Base decoder (H=8): A row of books sitting on a wooden bench.</p> <p>LSTM decoder (H=8): A wooden bench sitting on the side of a wooden bench.</p> <p>AoA decoder (H=8): A row of benches sitting on a wooden bench.</p> <p>Full (AoANet): A book sitting on top of a wooden bench.</p>	<p>GT1: A book on a wooden bench on a street.</p> <p>GT2: A book rest on one of the benches.</p> <p>GT3: A california travel guide book on a park bench.</p> <p>GT4: A big entitled california is on the bench.</p> <p>GT5: A book sitting on top of a wooden desk.</p>
	<p>Base model: A mountain with a mountain in the background.</p> <p>Encoder (w/ refining, w/o AoA): A view of a mountain with a mountain in the background.</p> <p>Encoder (w/ refining, w/ AoA): A black and white photo of a mountain.</p> <p>LSTM decoder (H=1): A mountain in the background with a mountain.</p> <p>AoA decoder (H=1): A black and white photo of a mountain on a mountain.</p> <p>Base decoder (H=8): A view of a mountain with a mountain.</p> <p>LSTM decoder (H=8): A mountain with a mountain in the background.</p> <p>AoA decoder (H=8): A view of a mountain with a mountain in the background.</p> <p>Full (AoANet): A black and white photo of a mountain covered with snow.</p>	<p>GT1: A black and white picture of hills or mountains.</p> <p>GT2: A black and white photo of mountains and the sky.</p> <p>GT3: A black and white photograph of mountains with snow.</p> <p>GT4: A black and white picture of the top view of a snow capped mountain somewhere.</p> <p>GT5: A large mountain range partially covered with snow.</p>
	<p>Base model: A brown bear eating a piece of food.</p> <p>Encoder (w/ refining, w/o AoA): A brown bear eating a piece of food.</p> <p>Encoder (w/ refining, w/ AoA): A monkey eating a doughnut in a rock.</p> <p>LSTM decoder (H=1): A brown bear sitting on top of a rock eating a piece of food.</p> <p>AoA decoder (H=1): A brown bear sitting on top of a rock eating a snack.</p> <p>Base decoder (H=8): A brown bear eating a banana in its mouth.</p> <p>LSTM decoder (H=8): A brown bear eating a piece of food.</p> <p>AoA decoder (H=8): A brown bear eating a piece of food on a rock.</p> <p>Full (AoANet): A monkey sitting on rocks eating a piece of food.</p>	<p>GT1: Small monkey eating fruit sitting on a rock.</p> <p>GT2: A small monkey sitting and eating in its habitat.</p> <p>GT3: A little monkey sitting next to a rock eating a piece of fruit.</p> <p>GT4: A monkey is eating while sitting on a rock.</p> <p>GT5: A monkey sits downs while eating some food.</p>
	<p>Base model: A vase with purple flowers sitting on a table.</p> <p>Encoder (w/ refining, w/o AoA): A vase with flowers sitting on top of a table.</p> <p>Encoder (w/ refining, w/ AoA): A white vase with purple flowers sitting on a table.</p> <p>LSTM decoder (H=1): A vase with flowers sitting on top of a table.</p> <p>AoA decoder (H=1): A vase filled with purple flowers sitting on a table.</p> <p>Base decoder (H=8): A white vase with purple flowers sitting on a table.</p> <p>LSTM decoder (H=8): A vase with flowers sitting on a table.</p> <p>AoA decoder (H=8): A white vase filled with purple flowers on a table.</p> <p>Full (AoANet): A white vase filled with purple flowers on a table.</p>	<p>GT1: A vase of light and dark purple flowers in a white vase.</p> <p>GT2: A close up of a vase with many flowers.</p> <p>GT3: A bouquet of fresh flowers in a cream colored vase.</p> <p>GT4: A small white vase of purple flowers on a table.</p> <p>GT5: A white vase filled with purple flowers on top of a table.</p>
	<p>Base model: A cat laying on top of a bed.</p> <p>Encoder (w/ refining, w/o AoA): A black and white cat laying on a bed.</p> <p>Encoder (w/ refining, w/ AoA): A cat laying on top of a blanket.</p> <p>LSTM decoder (H=1): A cat laying on top of a bed.</p> <p>AoA decoder (H=1): A cat laying on top of a blanket.</p> <p>Base decoder (H=8): A cat laying on top of a bed.</p> <p>LSTM decoder (H=8): A cat laying on top of a bed.</p> <p>AoA decoder (H=8): A cat laying on top of a blanket.</p> <p>Full (AoANet): A cat laying on a bed with a blanket.</p>	<p>GT1: A cat under a blanket looking at something.</p> <p>GT2: A small cat peeking its head from under a blanket.</p> <p>GT3: A cute kitten peeking out from under a quilt.</p> <p>GT4: A gray tiger cat sleeping on a bed under a blanket.</p> <p>GT5: A cat is peeping out from under a blanket.</p>
	<p>Base model: A traffic light on the side of a street.</p> <p>Encoder (w/ refining, w/o AoA): A traffic light on the side of a street.</p> <p>Encoder (w/ refining, w/ AoA): A group of traffic lights on a city street.</p> <p>LSTM decoder (H=1): A traffic light on a city street with a building.</p> <p>AoA decoder (H=1): A traffic lights on a city street with buildings.</p> <p>Base decoder (H=8): A traffic light on the side of a city street.</p> <p>LSTM decoder (H=8): A traffic light on the side of a street.</p> <p>AoA decoder (H=8): A traffic light on a city street with a city.</p> <p>Full (AoANet): A street with traffic lights and buildings in the background.</p>	<p>GT1: A couple of traffic lights hanging over a city street.</p> <p>GT2: Traffic lights shine over an empty intersection at twilight.</p> <p>GT3: A street with traffic lights, wall and buildings.</p> <p>GT4: The electronic traffic signals are lit up during dawn.</p> <p>GT5: Empty city intersection in the early morning hours.</p>
	<p>Base model: A brown horse standing in the water next to a lake.</p> <p>Encoder (w/ refining, w/o AoA): Two horses are standing in the water.</p> <p>Encoder (w/ refining, w/ AoA): Two horses are standing in the water.</p> <p>LSTM decoder (H=1): A horse standing in the water in a field.</p> <p>AoA decoder (H=1): A horse standing in the water near a river.</p> <p>Base decoder (H=8): A brown horse standing in the water.</p> <p>LSTM decoder (H=8): A horse standing in the water in a field.</p> <p>AoA decoder (H=8): A horse standing in the water next to a river.</p> <p>Full (AoANet): Two horses walking in the water in a field.</p>	<p>GT1: A couple of horses standing in a river next to an island.</p> <p>GT2: Two horses standing in a stream next to a wooded area.</p> <p>GT3: Two horses are walking through a river together.</p> <p>GT4: Horses standing in shallow water in a wooded area.</p> <p>GT5: Two horses that are standing in some water.</p>

Figure 1: Captions generated by the ablated models and AoANet, as well as the corresponding ground truths.

Image	Generated captions	Ground truths
	<p>Base model: A man standing in front of a car with a UNK.</p> <p>Encoder (w/ refining, w/o AoA): A man wearing a hat standing next to a beach with a frisbee.</p> <p>Encoder (w/ refining, w/ AoA): A man standing in a parking lot with a frisbee.</p> <p>LSTM decoder (H=1): A man is standing next to a car on a water.</p> <p>AoA decoder (H=1): A man is standing on a car with a frisbee.</p> <p>Base decoder (H=8): A man standing in a parking lot with a frisbee.</p> <p>LSTM decoder (H=8): A man sitting on a parking lot with a boat.</p> <p>AoA decoder (H=8): A man standing in a parking lot with a frisbee.</p> <p>Full (AoANet): A man wearing a hat holding a frisbee in a parking lot.</p>	<p>GT1: A man in holds a frisbee in the parking lot by the river.</p> <p>GT2: The man found the frisbee by the parked cars.</p> <p>GT3: A man holding a frisbee in a parking lot near water.</p> <p>GT4: A man holding a white object in his hand.</p> <p>GT5: A man wearing a hat and holding an object in his hand.</p>
	<p>Base model: A pool with a pool with people and umbrellas.</p> <p>Encoder (w/ refining, w/o AoA): A group of people standing in a park with umbrellas.</p> <p>Encoder (w/ refining, w/ AoA): A group of people standing in a pool with umbrellas.</p> <p>LSTM decoder (H=1): A pool with a pool with a pool with a pool.</p> <p>AoA decoder (H=1): A swimming pool with a pool with people.</p> <p>Base decoder (H=8): A pool with a pool with umbrellas on a pool.</p> <p>LSTM decoder (H=8): A swimming pool with a bunch of chairs and a table.</p> <p>AoA decoder (H=8): A swimming pool with tables and chairs with umbrellas.</p> <p>Full (AoANet): A swimming pool with umbrellas and people standing.</p>	<p>GT1: The clear pool stood empty as people stood around nearby.</p> <p>GT2: Some people walking around a swimming pool and trees.</p> <p>GT3: Large swimming pool with tents near the trees.</p> <p>GT4: A large in ground swimming pool near tents.</p> <p>GT5: A pool with several visitors sitting beside it.</p>
	<p>Base model: A group of people walking down a street with skateboards.</p> <p>Encoder (w/ refining, w/o AoA): A man and a child riding on a skateboard.</p> <p>Encoder (w/ refining, w/ AoA): A man and a little girl on a skateboard.</p> <p>LSTM decoder (H=1): A group of people riding skateboards down a street.</p> <p>AoA decoder (H=1): A man and a child on a skateboard in the street.</p> <p>Base decoder (H=8): A group of people riding skateboards down a street.</p> <p>LSTM decoder (H=8): A man and a woman riding a skateboard down a street.</p> <p>AoA decoder (H=8): A man and a child on a skateboard.</p> <p>Full (AoANet): A man and a child sitting on a skateboard.</p>	<p>GT1: A girl hitching a ride with her dad on a skateboard.</p> <p>GT2: A man riding a skateboard has a child between his legs.</p> <p>GT3: A girl sits between a mans legs on a skateboard.</p> <p>GT4: A man riding a skateboard next to a woman.</p> <p>GT5: The skateboarder has a passenger on his skateboard.</p>
	<p>Base model: A man and a woman standing in a kitchen preparing food.</p> <p>Encoder (w/ refining, w/o AoA): A group of people standing in a kitchen.</p> <p>Encoder (w/ refining, w/ AoA): A group of people standing in a kitchen preparing food.</p> <p>LSTM decoder (H=1): A man and a woman standing in a kitchen.</p> <p>AoA decoder (H=1): A group of people standing in a kitchen.</p> <p>Base decoder (H=8): A man and a woman standing in a kitchen.</p> <p>LSTM decoder (H=8): A man and a woman standing in a kitchen.</p> <p>AoA decoder (H=8): A man and a woman standing in a kitchen preparing food.</p> <p>Full (AoANet): A group of people standing in a kitchen preparing food.</p>	<p>GT1: A picture of an older man cutting a turkey on the counter.</p> <p>GT2: A group of people that are standing in the kitchen.</p> <p>GT3: A group of people standing over some food.</p> <p>GT4: Three people agree in the kitchen near a cluttered bar.</p> <p>GT5: A number of people in a kitchen preparing food.</p>
	<p>Base model: A picture of a garden with a plant in a window.</p> <p>Encoder (w/ refining, w/o AoA): A bunch of apples in a window with a window.</p> <p>Encoder (w/ refining, w/ AoA): A garden with a bunch of apples and a window.</p> <p>LSTM decoder (H=1): A picture of a bunch of plants in a house.</p> <p>AoA decoder (H=1): A collage of pictures of a house with a tree.</p> <p>Base decoder (H=8): A group of plants in a window with a table.</p> <p>LSTM decoder (H=8): A picture of a tree with a bunch of apples.</p> <p>AoA decoder (H=8): A collage of pictures of a garden with a tree.</p> <p>Full (AoANet): A collage of pictures of a house with a flower.</p>	<p>GT1: Multiple images are seen showing forest scenes and a small cottage.</p> <p>GT2: A collection of different plants and fields, with a tomato plant being the most prominent.</p> <p>GT3: Various types of pictures that are of different plants.</p> <p>GT4: Colorful collection of pictures of a wooded area.</p> <p>GT5: A collage of several photos show outdoor scenes.</p>
	<p>Base model: A man hitting a tennis ball with a tennis racket.</p> <p>Encoder (w/ refining, w/o AoA): A group of people playing tennis on a tennis court.</p> <p>Encoder (w/ refining, w/ AoA): A group of people playing tennis on a tennis court.</p> <p>LSTM decoder (H=1): A group of people playing tennis on a tennis court.</p> <p>AoA decoder (H=1): A couple of women standing on a tennis court with tennis balls.</p> <p>Base decoder (H=8): A couple of people holding tennis rackets at a tennis ball.</p> <p>LSTM decoder (H=8): A woman hitting a tennis ball with a tennis racket.</p> <p>AoA decoder (H=8): Two women playing tennis on a tennis court.</p> <p>Full (AoANet): Two people standing on a tennis court holding tennis rackets.</p>	<p>GT1: Two young sexy women holding tennis racquets and tennis balls.</p> <p>GT2: A pair of young women hold tennis balls and rackets.</p> <p>GT3: Two women on a tennis court with tennis rackets and balls.</p> <p>GT4: Two female tennis players standing with their rackets.</p> <p>GT5: Two women with tennis rackets and balls are standing.</p>

Figure 2: More examples.