Video-LLaVA (**Video Captioning**)

NLTK (**Instance Highlight**)

BLIP2 (**Image Captioning**)

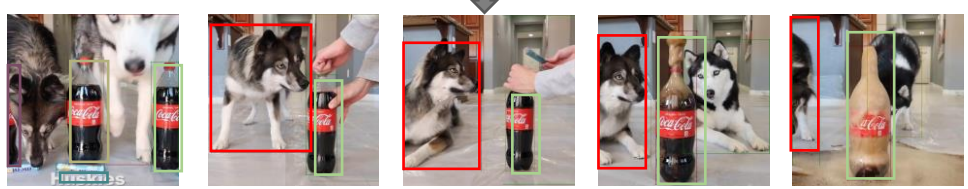
a man with glasses and a dog in front of a barn	a man with glasses is taking a picture of a dog	a man in glasses is petting a doberman	a man in glasses is standing next to a dog	a man with glasses and a dog standing next to him

The video features two **canines**, one black and white and the other brown and white, sitting on the floor and looking at a **beverage**. The black and white **canine** sniffs the bottle and then licks it, while the brown and white **canine** also sniffs the bottle and licks it. The video ends with the two **canines** sitting on the floor.

Prompt: ***"Enhance continuity and storytelling in captions…"*** Large Language Model (**Object Alignment**)

canine are positioned beside the bottled **beverage**	a **human** is gripping a bottled **beverage** while **canine** lingers in the distance	a human is clasping a bottled **beverage** a **canine** take a seat on it	**canine** are lounging next to a **beverage**	**canine** are frolicking with a **beverage**

YOLO-World (**Object Detection**)



DINO (**Image-Text Entity Alignment**)

EfficientViT-SAM (**Instance Segmentation**)

Aligned caption list

canine0 and **canine1** are positioned beside the bottled **beverage0** and **beverage1** and **beverage2**	a human is gripping a bottled **beverage0** while **canine0** lingers in the distance	a human is clasping a bottled **beverage0** a **canine0** take a seat on it	**canine0** and **canine1** are lounging next to a **beverage0**	**canine0** and **canine1** are frolicking with a **beverage0**

Target frame sequence