

A. Additional training details

We implemented CaptionT5 using PyTorch [24], Hugging-Face Transformers [36], and OpenAI CLIP [26]. Following VALUE [15], we used the AdamW [21] optimizer with a linear learning rate scheduling. We trained CaptionT5 on 16 NVIDIA A100 GPUs. Hyperparameters used for training CaptionT5 are summarized in Table 8.

We implemented CaptionT5 to choose one of two video encoding methods during training. One option is to encode raw video frames by using the CLIP image encoder on-the-fly. The other option is to use offline image features encoded by the CLIP image encoder before the training. Even though the first option takes longer training time, it can benefit from using better image features encoded by a fine-grained CLIP image encoder such as CLIP-ViT-B/16. The second option has the advantage of fast training, but we may not take advantage of using better and diverse image features. In the experiment, when training CaptionT5 on VA-TEX [34] dataset and YC2 [46], we used offline CLIP-ViT-B/32 image features provided by the VALUE [15] benchmark. Therefore, we could iterate many experiments leveraging fast training. On the other hand, for learning on MSR-VTT [38] dataset, we used online image features encoded by using CLIP-ViT-B/16.

training, but they are retrieved by top-k search for inference to generate deterministic results.

Hyperparameter	Value
Optimizer	AdamW [21]
Betas	(0.9, 0.999)
Weight decay	0.1
Adam epsilon	1e-8
LR scheduler	Linear
Initial LR	1.5e-4
Warm-up steps	5000
Batch size	256
Max epochs	20
Temperature for TS	1e-3
Beam search length (default)	6
Beam search length for CR	2~10

Table 8. **Hyperparameters used for training CaptionT5.** LR, TS, and CR means learning rate, thought sampling, and caption ranking, respectively.

B. Additional qualitative results

We provide additional qualitative results in Figure 5, 6, and 7. The results are generated by CaptionT5 trained in the setting where 8 video frames are given, and 7 object prompts and 7 action prompts are retrieved. Note that thought prompts are retrieved by similarity sampling for

		
VideoID		EA3HCx0yTIY_000281_000291
CaptionT5	Object prompts	A photo of microphone, mike. A photo of tick. A photo of jay. A photo of drumstick. A photo of steel drum. A photo of poncho. A photo of crutch.
	Action prompts	A photo of playing drums. A photo of drumming fingers. A photo of playing cymbals. A photo of playing gong. A photo of air drumming. A photo of recording music. A photo of finger snapping.
	Generated caption	A man is playing a set of drums and cymbals.
SwinBERT		A man is sitting at a drum set and playing the drums.
GT1		A man sits and plays music on a set of drums.
GT2		A bearded, bald drummer sits at a drumset and strikes a cymbal five times.
		
VideoID		G0mjFqytJt4_000152_000162
CaptionT5	Object prompts	A photo of envelope. A photo of carton. A photo of sombrero. A photo of quill, quill pen. A photo of sarong. A photo of wing. A photo of pole.
	Action prompts	A photo of making paper aeroplanes. A photo of ripping paper. A photo of applying cream. A photo of poking bellybutton. A photo of tapping pen. A photo of pinching. A photo of beatboxing.
	Generated caption	A young boy is demonstrating how to fold a paper airplane.
SwinBERT		A young boy is showing how to make a paper airplane
GT1		A boy is talking and fiddling with a few pieces of paper in his hands.
GT2		A young boy in his bathroom as he explains how to make a paper airplane.
		
VideoID		IczD9OzKvco_000102_000112
CaptionT5	Object prompts	A photo of rocking chair, rocker. A photo of folding chair. A photo of crib, cot. A photo of mailbag, postbag. A photo of cradle. A photo of bib. A photo of abacus.
	Action prompts	A photo of crawling baby. A photo of moving baby. A photo of moving child. A photo of shouting. A photo of falling off chair. A photo of air drumming. A photo of clapping.
	Generated caption	A baby is sitting in a high chair and shaking his head back and forth.
SwinBERT		A baby is sitting in a high chair and shaking his head back and forth.
GT1		A little boy sitting in a high chair is cooing and shaking his head, while a woman is talking to him and shaking her head.
GT2		A baby sits in a highchair, shakes his head and smiles as an adult watches and laughs.
		
VideoID		Pj_070vBUeQ_000010_000020
CaptionT5	Object prompts	A photo of barbell. A photo of balance beam, beam. A photo of dumbbell. A photo of punching bag, punch bag, punching ball, punchball. A photo of parallel bars, bars. A photo of horizontal bar, high bar. A photo of knee pad.
	Action prompts	A photo of rope pushdown. A photo of jumping jacks. A photo of clean and jerk. A photo of snatch weight lifting. A photo of deadlifting. A photo of exercising with an exercise ball. A photo of push up.
	Generated caption	A man is teaching a woman how to do lunges in a gym.
SwinBERT		A man and a woman are doing jumping jacks in a gym.
GT1		A man and a woman are making an instructional video on the proper way to do jumping jacks.
GT2		A man demonstrates how to work out with a woman doing the work outs.

Figure 5. Example captions generated by CaptionT5 on VATEX dataset.

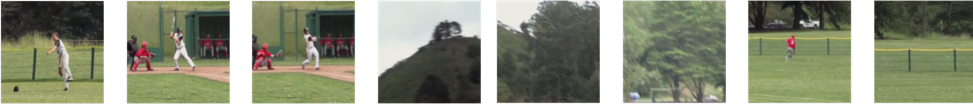
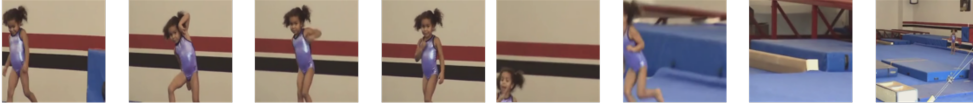


		
VideoID		video7021
CaptionT5	Object prompts	A photo of baseball. A photo of ballplayer, baseball player. A photo of pole. A photo of swing. A photo of pitcher, ewer. A photo of projectile, missile. A photo of cliff, drop, drop-off.
	Action prompts	A photo of catching or throwing baseball. A photo of hitting baseball. A photo of catching or throwing softball. A photo of throwing ball (not baseball or American football). A photo of throwing axe. A photo of swinging baseball bat. A photo of catching or throwing frisbee.
	Generated caption	A baseball player is hitting the ball with his bat.
GT1		Baseball player hits ball.
		
VideoID		video9771
CaptionT5	Object prompts	A photo of balance beam, beam. A photo of parallel bars, bars. A photo of horizontal bar, high bar. A photo of bow. A photo of ski. A photo of fly. A photo of velvet.
	Action prompts	A photo of gymnastics tumbling. A photo of jumpstyle dancing. A photo of shoot dance. A photo of jumping jacks. A photo of moving child. A photo of drop kicking. A photo of high kick.
	Generated caption	A girl is doing gymnastics.
GT1		A little girl does gymnastics.
		
VideoID		video9773
CaptionT5	Object prompts	A photo of pickup, pickup truck. A photo of sports car, sport car. A photo of scorpion. A photo of moped. A photo of moving van. A photo of slot, one-armed bandit. A photo of torch.
	Action prompts	A photo of pushing car. A photo of directing traffic. A photo of falling off bike. A photo of jaywalking. A photo of riding a bike. A photo of driving car. A photo of motorcycling.
	Generated caption	A person is playing a video game.
GT1		A boy plays Grand Theft Auto 5.
		
VideoID		video9779
CaptionT5	Object prompts	A photo of pirate, pirate ship. A photo of fireboat. A photo of torch. A photo of maypole. A photo of pole. A photo of parachute, chute. A photo of crane.
	Action prompts	A photo of shooting off fireworks. A photo of celebrating. A photo of card throwing. A photo of lighting fire. A photo of sailing. A photo of popping balloons. A photo of breathing fire.
	Generated caption	Fireworks are exploding in the sky.
GT1		Fireworks are being lit and exploding in a night sky.

Figure 6. Example captions generated by CaptionT5 on MSR-VTT dataset.





		
VideoID		efnHOsT7k9s_6
CaptionT5	Object prompts	A photo of pizza, pizza pie. A photo of dough. A photo of frying pan, frypan, skillet. A photo of zucchini, courgette. A photo of plate. A photo of pinwheel. A photo of pineapple, ananas.
	Action prompts	A photo of making pizza. A photo of slicing onion. A photo of blending fruit. A photo of cutting pineapple. A photo of cutting apple. A photo of frying vegetables. A photo of preparing salad.
	Generated caption	Place basil leaves on top of the pizza
SwinBERT		Place the basil on the pizza
GT		Place basil leaves on top of the pizza
		
VideoID		tYg3lQ5aZv8_2
CaptionT5	Object prompts	A photo of cucumber, cuke. A photo of broccoli. A photo of guacamole. A photo of zucchini, courgette. A photo of syringe. A photo of spatula. A photo of iron, smoothing iron.
	Action prompts	A photo of slicing onion. A photo of preparing salad. A photo of frying vegetables. A photo of using a paint roller. A photo of cutting apple. A photo of making slime. A photo of making sushi.
	Generated caption	Chop green onions and add them to the bowl
SwinBERT		Chop the green onion
GT		Finely chop green onions
		
VideoID		E9O9-6TQUw0_2
CaptionT5	Object prompts	A photo of cleaver, meat cleaver, chopper. A photo of butcher shop, meat market. A photo of meat loaf, meatloaf. A photo of hot pot, hotpot. A photo of frying pan, frypan, skillet. A photo of harvester, reaper. A photo of spatula.
	Action prompts	A photo of chopping meat. A photo of grinding meat. A photo of cutting apple. A photo of poaching eggs. A photo of carving ice. A photo of waxing chest. A photo of cutting watermelon.
	Generated caption	Add pepper to the meat
SwinBERT		Season the meat with salt and pepper
GT		Sprinkle salt and pepper on top of the meat
		
VideoID		0uaKitJaqml_7
CaptionT5	Object prompts	A photo of mixing bowl. A photo of frying pan, frypan, skillet. A photo of ladle. A photo of mashed potato. A photo of wok. A photo of Dutch oven. A photo of dough.
	Action prompts	A photo of peeling potatoes. A photo of cooking egg. A photo of cooking scallops. A photo of slicing onion. A photo of making a cake. A photo of shaping bread dough. A photo of separating eggs.
	Generated caption	Mesh the potatoes with a potato masher
SwinBERT		Mash the potatoes with salt
GT		Mash the potatoes

Figure 7. Example captions generated by CaptionT5 on YC2 dataset.