COOT: Cooperative Hierarchical Transformer for Video-Text Representation Learning

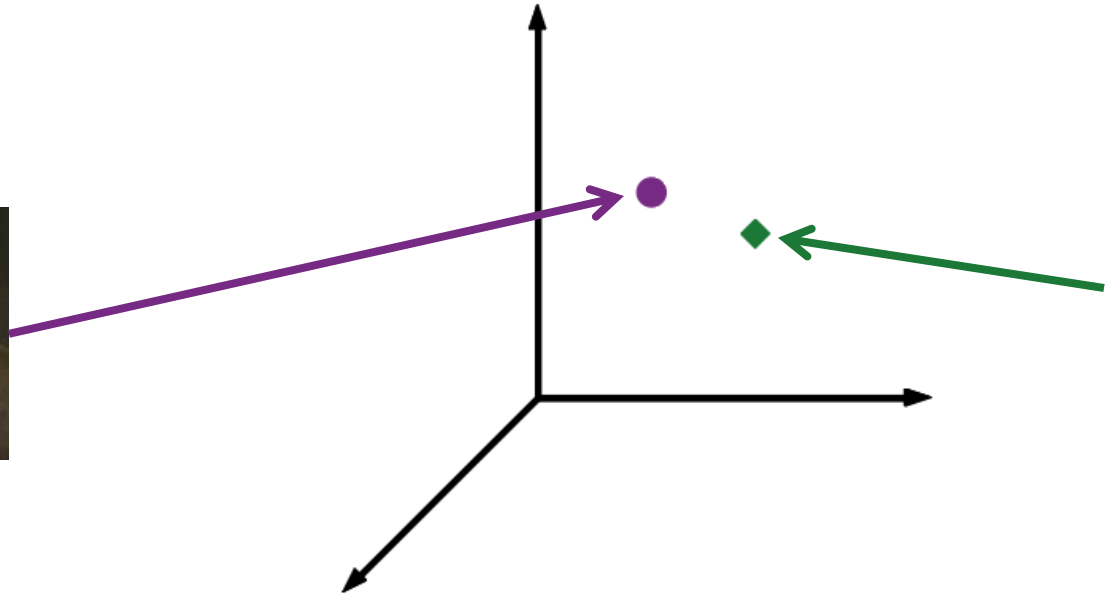
Simon Ging[1*], **Mohammadreza Zolfaghari**[1*], **Hamed Pirsiavash**[2], **Thomas Brox**[1]

[1] University of Freiburg, [2] University of Maryland Baltimore County

[1]{gings, zolfagha, brox}@cs.uni-freiburg.de, [2]hpirsiav@umbc.edu

[*] Equal contribution

Video-Text Representation Learning



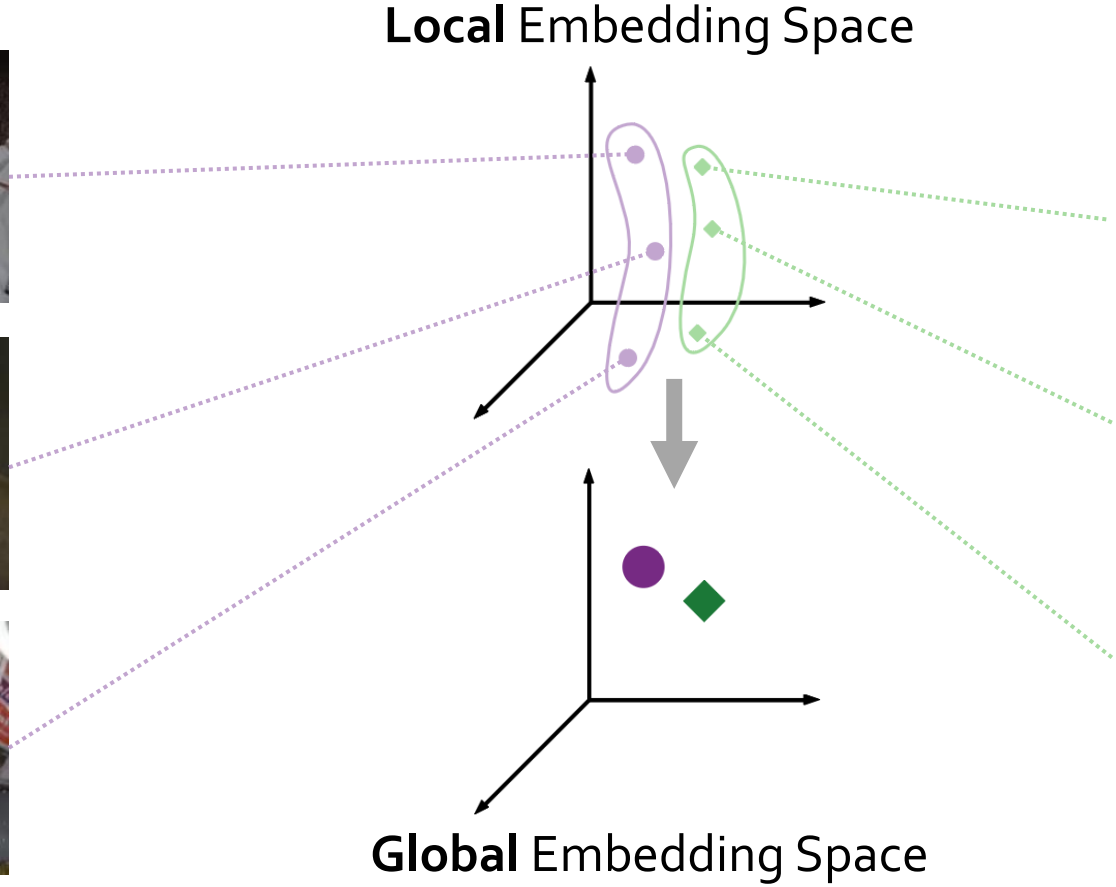
A white yard chair is being shown.

A man scrapes the paint off the chair.

He uses the white paint to coat the chair.

Embedding Space

Video-Text Representation Learning

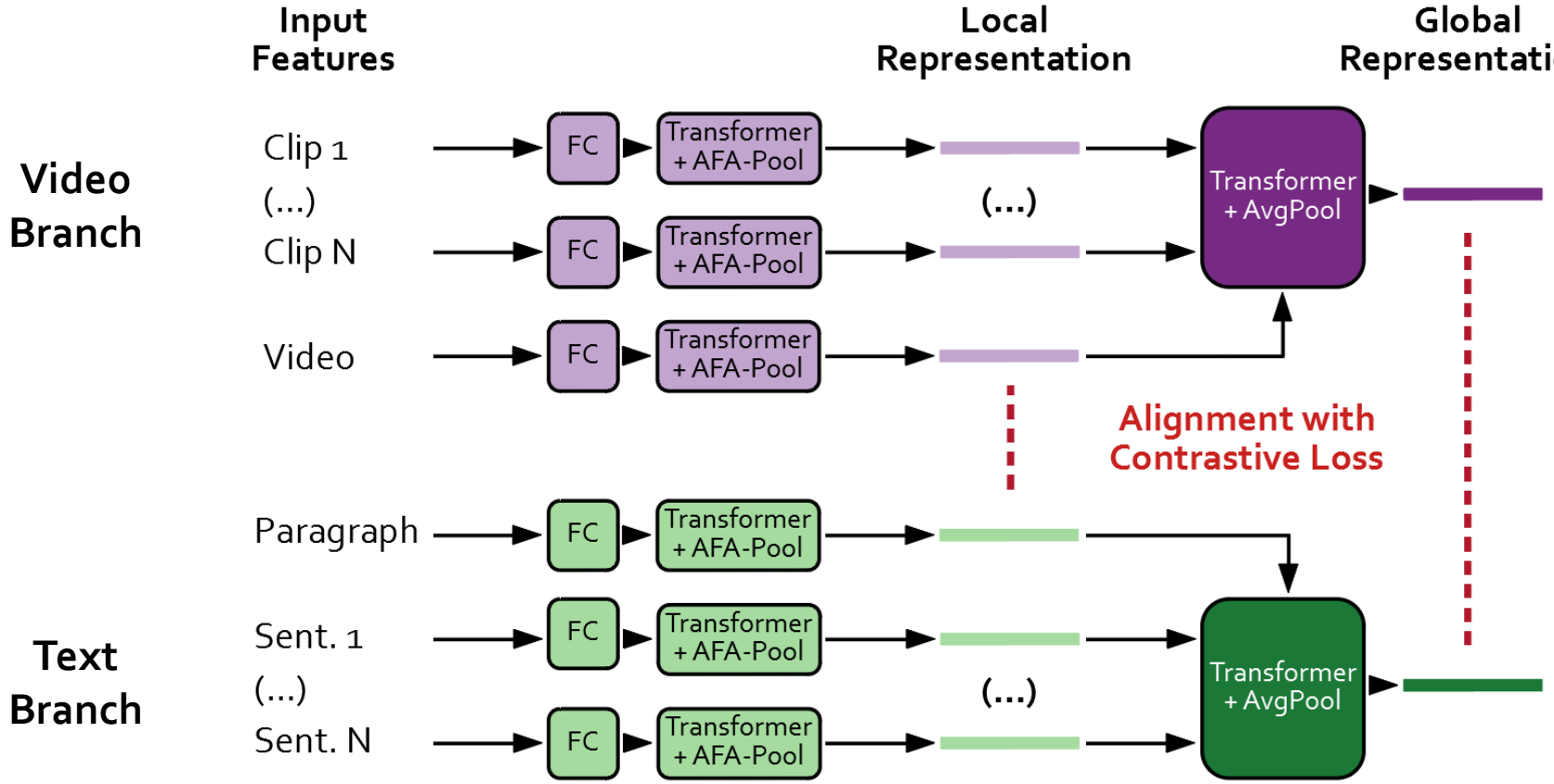


A white yard chair is being shown.

A man scrapes the paint off the chair.

He uses the white paint to coat the chair.

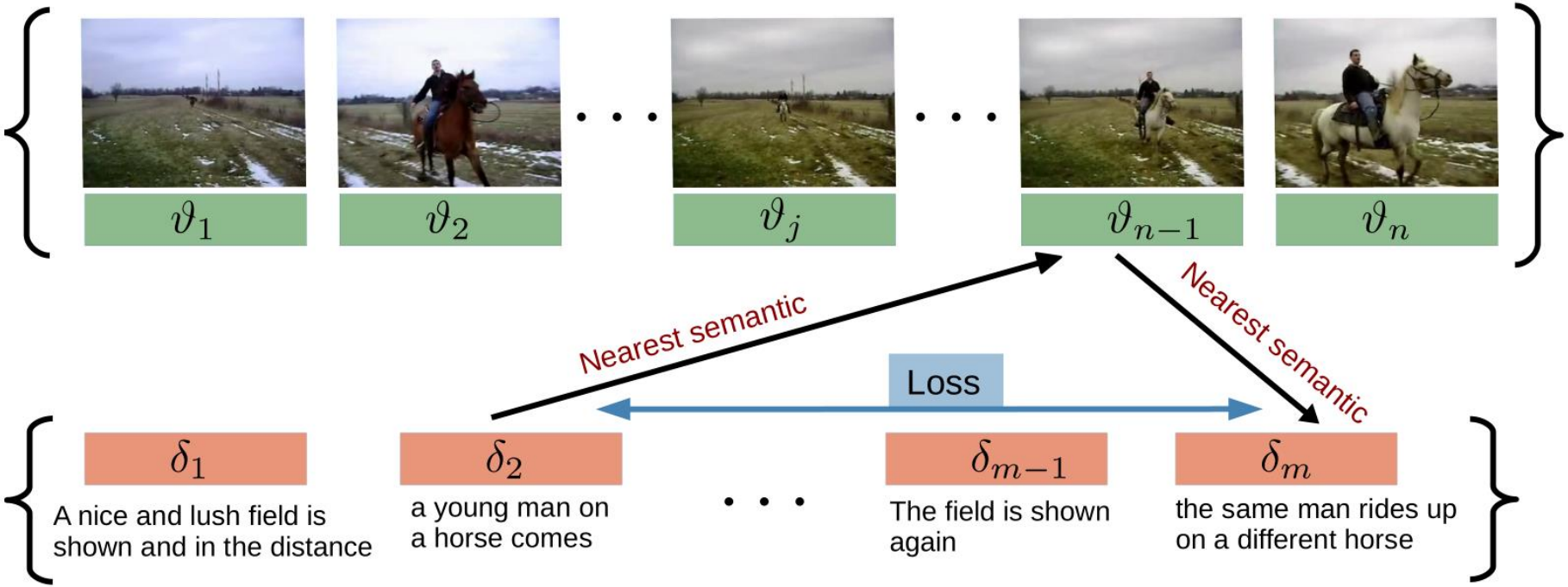
COOT: Cooperative Hierarchical Transformer



AFA-Pool: Attention-aware Feature Aggregation

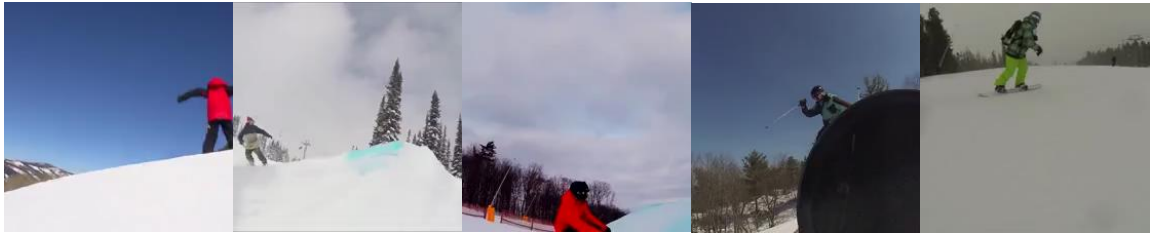
Cross-Modal Cycle Consistency

- Use **Temporal Ordering** to enforce cross-model interaction
- **Cycle** between Neighbours in Clip-Sentence Space



Retrieval Results

Query: A young person goes snowboarding on a mountain. He goes down, jumps and flips and spins.



ActivityNet		Paragraph ⟹ Video			
Method	Published	R@1	R@5	R@50	MR
CMHSE	ECCV (2018)	44.4±0.5	76.7±0.3	97.1±0.1	2
COOT (Ours)	NeurIPS (2020)	**60.8**±0.6	**86.6**±0.4	**98.6**±0.1	**1**

Youcook2		Par. ⟹ Video		Sent. ⟹ Clip	
Method	Published	R@1	R@5	R@1	R@5
Miech et al.	ICCV (2019)	59.6	86.0	8.2	24.5
ActBert	CVPR (2020)	-	-	9.6	26.7
MIL-NCE	CVPR (2020)	61.9	89.4	15.1	38.0
COOT (Ours)	NeurIPS (2020)	**77.2**±1.0	**95.8**±0.8	**16.7**±0.4	**40.2**±0.3

Captioning Results

- Use MART method (Transformer-based Captioning)
- Replace Appearance + Flow features with COOT embeddings
  - We input ~100 times less video data into the Captioning Model

ActivityNet

Method	Published	BLEU@4	RougeL	METEOR	CIDEr-D
MART	ACL (2020)	9.78	30.63	15.57	22.16
COOT (Ours)	NeurIPS (2020)	**10.85**	**31.45**	**15.99**	**28.19**

Youcook2

Method	Published	BLEU@4	RougeL	METEOR	CIDEr-D
MART	ACL (2020)	8.00	31.97	15.90	35.74
COOT (Ours)	NeurIPS (2020)	**11.30**	**37.94**	**19.85**	**57.24**

Captioning Results (YouCook2 *val*)

Ground Truth: Chop *celery*, *apple*, *red grapes* and *roasted walnuts*. Whisk *mayonnaise*, *lemon juice* and *pepper* and combine with the *fruits* and *nuts*. Place the *salad* on *lettuce*.

MART, ACL (2020): Chop some *red onion and garlic*. Add the *beef* to a bowl and mix. Add *cabbage* and *cabbage* to the bowl.

COOT (Ours): Chop the *celery root* and add them to a bowl. Add *lemon juice*, *olive oil*, *salt* and *pepper* to the bowl and mix well. Toss the *salad*.

Captioning Results (ActivityNet Captions *ae-test*)

Ground Truth: A *group of people* is preparing for a *rafting* trip. A *GoPro* is used to capture parts of the trip.

MART, ACL (2020): People are rafting down a river. They are rafting down a river.

COOT (Ours): A large group of people are seen sitting in a raft and leads into them riding down a river. The people continue riding down the river past one another and end by holding a selfie stick up.

Thank you for your attention!