Connecting Vision and Language with Video Localized Narratives

Paul Voigtlaender, Soravit Changpinyo, Jordi Pont-Tuset, Radu Soricut, Vittorio Ferrari

TUE-AM-235

Google Research

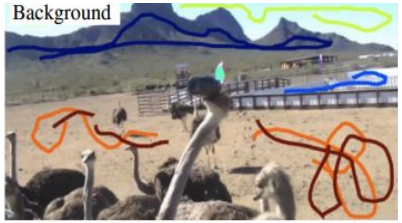
Video Localized Narratives

- Multi-modal video descriptions
- Localized in space-time
- Capture the story of the video
- 50k videos annotated for OVIS, UVO, Oops, and Kinetics
- Defined two tasks on the data
  - Video Narrative Grounding
  - Video Question Answering

<Man> A man wearing a black t-shirt is holding a cup of food in his right hand. He moves around a piece of food in his left hand to play with the ostrich.

<Ostrich> An ostrich is looking at the piece of food held by the man and suddenly grabs the cup of food and starts eating.

<Background> In the background, there are hills, white barriers, a flag, the sky, and soil on the ground.



Annotation Process

- Annotators talk while moving their mouse over keyframes
- Speech and mouse pointer synchronized → localize words with trace segments

- Key-frames avoid “race against time”
- Focus on the story of the video, from the viewpoint of each actor
- Mention actor name, its attributes, actions it performs on other actors (ostrich) and on passive objects (cup)
- One narration per actor → disentangle situations → capture complex events

Statistics: 4 datasets, 50k videos, 3.5M words

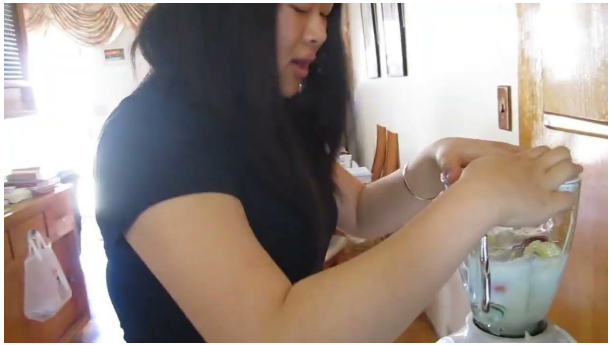
OVIS - 607 videos

[Qi IJCV'22]

Kinetics

30k videos

[Kay arXiv'17]

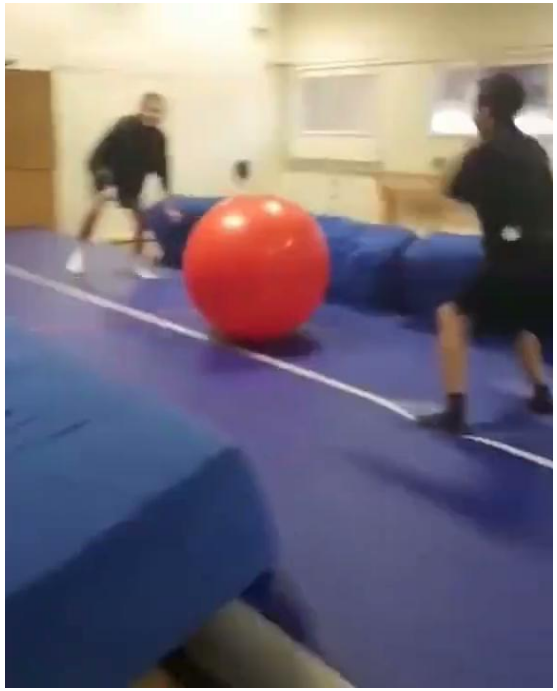


UVO - 8.6k videos

[Wang ICCV'21]

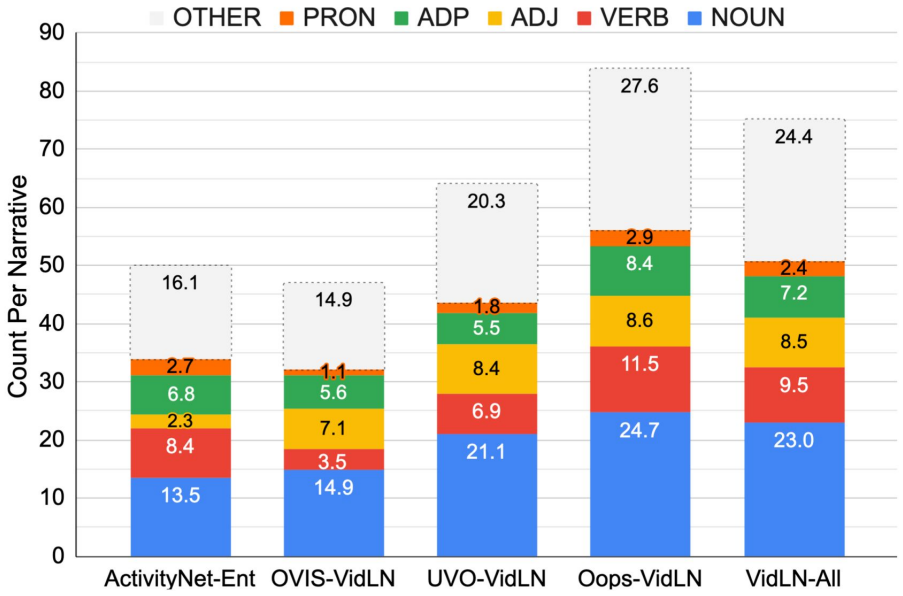
Oops- 12k videos

[Epstein CVPR'20]



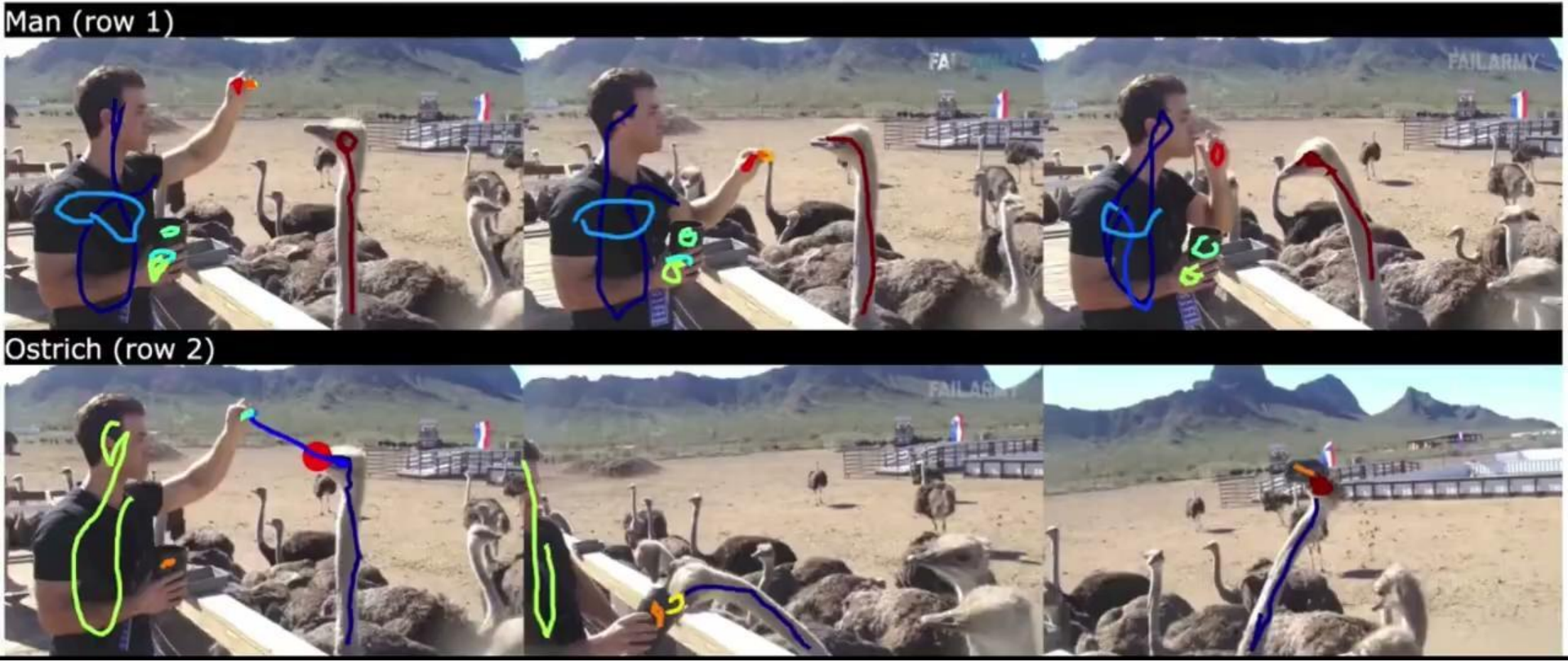
Rich Annotations

- Long captions: ~75 words per narrative
- ~97% semantic accuracy, 73%-93% mouse trace precision (depending on setting, see paper)
- More nouns, verbs, adjectives, adpositions than closest dataset (ActivityNet)



Analysis on OVIS, UVO, Oops

<Man> A man wearing a black t-shirt is holding a cup of food in his right hand. He moves around a piece of food in his left hand to play with the ostrich.
<Ostrich> An ostrich is looking at the piece of food held by the man and suddenly grabs the cup of food and starts eating.



Task 1: Video Narrative Grounding (VNG)

- Task: Localize each noun with a segmentation mask in each frame
- Multiple identical nouns need to be disambiguated using context provided by other words

VNG method

predict

[Person]: A person is holding the parrots in hand.

[Parrot one]: A green parrot with a red-black neckline is playing with the other parrot.

[Parrot two]: Another green-red parrot is sitting on person's hand.

Task 1: Video Narrative Grounding

- On video segmentation datasets OVIS and UVO (8k videos with 45k nouns)
- We connected nouns to existing masks and added extra masks

- Related task Ref-VOS [Seo ECCV'20]: 1 sentence = 1 object
- VNG more challenging, need to localize multiple nouns, disambiguate using context by other words
- 2x more videos and 3x more objects than Ref-VOS datasets

Benchmark	Task	Videos	Objects (per vid.)
OVIS-VNG	VNG	505	2,407 (4.77)
UVO-VNG	VNG	7,587	43,058 (5.68)
Ref-YTB-VOS [47]	R-VOS	3,978	15,009 (3.77)
Ref-DAVIS'17 [22]	R-VOS	90	205 (2.28)

- ReferFormer-VNG baseline accuracy: 32.7% OVIS / 46.4% UVO

Task 2: Video Question Answering

- Questions+answers automatically generated from narrative (with VQ²A method [Changpinyo NAACL'22])
- Verified by annotators
- Text-output questions (44k Qs on 9.5k videos)

  Q: "What is the shape of the girl's hat?"
  A: "cone"

  PaLI baseline accuracy: 49%

- Location-output questions (18k Qs on 9.8k videos)

  Ground truth traces have 93% accuracy

  Q: "Where is the girl that is wearing a pink dress?"
  A: mouse trace / bounding box

  ReferFormer baseline accuracy: 48.3%

Conclusion

- Video Localized Narratives for 50k videos
- New Video Narrative Grounding and Video Question Answering benchmarks
- Project page: https://google.github.io/video-localized-narratives/

<Man> A man wearing a black t-shirt is holding a cup of food in his right hand. He moves around a piece of food in his left hand to play with the ostrich.

<Ostrich> An ostrich is looking at the piece of food held by the man and suddenly grabs the cup of food and starts eating.