Grounded Video Description: Supplementary Document

Luowei Zhou[1,2], Yannis Kalantidis[1], Xinlei Chen[1], Jason J. Corso[2], Marcus Rohrbach[1]
[1] Facebook AI, [2] University of Michigan
github.com/facebookresearch/grounded-video-description

A. Appendix

This Appendix provides additional details, evaluations, and qualitative results.

- In Sec. A.1, we provide more details on our dataset including the annotation interface and examples of our dataset, which are shown in Figs. 1, 2.
- In Sec. A.2, we clarify on the four localization metrics.
- In Sec. A.3, we provide additional ablations and results on our ActivityNet-Entities dataset, including qualitative results, which are shown in Figs. 3, 4.
- In Sec. A.4, we provide additional results on the Flickr30kEntities dataset, including qualitative results, which are shown in Fig. 5.
- In Sec. A.5, we provide more implementation details (*e.g.*, training details).

A.1. Dataset

Definition of a noun phrase. Following the convention from Flickr30k Entities dataset [4], we define noun phrase as:

- short (avg. 2.23 words), non-recursive phrases (*e.g.*, the complex NP "the man in a white shirt with a heart" is split into three: "the man", "a white shirt", and "a heart")
- refer to a specific region in the image so as to be annotated as a bounding box.
- could be
  - a single instance (*e.g.*, a cat),
  - multiple distinct instances (*e.g.* two men),
  - a group of instances (*e.g.*, a group of people),
  - a region or scene (*e.g.*, grass/field/kitchen/town),
  - a pronoun, *e.g.*, it, him, they.
- could include
  - adjectives (*e.g.*, a *white* shirt),
  - determiners (*e.g.*, *A* piece of exercise equipment),
  - prepositions (*e.g.* the woman *on the right*)
  - other noun phrases, if they refer to the identical bounding concept & bounding box (*e.g.*, a group of people, a shirt of red color)

Annotator instructions
Further instructions include:

- Each word from the caption can appear in at most one NP. "A man in a white shirt" and "a white shirt" should not be annotated at the same time.
- Annotate multiple boxes for the same NP if the NP refers to multiple instances.
  - If there are more than 5 instances/boxes (*e.g.*, six cats or many young children), mark all instances as a single box and mark as "a group of objects".
  - Annotate 5 or fewer instances with a single box if the instances are difficult to separate, *e.g.* if they are strongly occluding each other.
- We don't annotate a NP if it's abstract or not presented in the scene (*e.g.*, "the camera" in "A man is speaking to the camera")
- One box can correspond to multiple NPs in the sentence (*e.g.*, "the man" and "him"), *i.e.*, we annotate co-references within one sentence.

See Fig. 1 for more examples.
Annotation interface. We show a screen shot of the interface in Fig. 2.
Validation process. We deployed a rigid quality control process during annotations. We were in daily contact with the annotators, encouraged them to flag all examples that were unclear and inspected a sample of the annotations

daily, providing them with feedback on possible spotted annotation errors or guideline violations. We also had a post-annotation verification process where all the annotations are verified by human annotators.

Dataset statistics. The average number of annotated boxes per video segment is 2.56 and the standard deviation is 2.04. The average number of object labels per box is 1.17 and the standard deviation is 0.47. The top ten frequent objects are "man", "he", "people", "they", "she", "woman", "girl", "person", "it", and "boy". Note that the statistics are on object boxes, *i.e.*, after pre-processing.

List of objects. Tab. 10 lists all the 432 object classes which we use in our approach. We threshold at 50 occurrences. Note that the annotations in ActivityNet-Entities also contain the full noun phrases w/o thresholds.

A.2. Localization Metrics

We use four localization metrics, $Attn.$, $Grd.$, $F1_{all}$, and $F1_{loc}$ as mentioned in Sec. 5.1. The first two are computed on the GT sentences, *i.e.*, during inference, we feed the GT sentences into the model and compute the attention and grounding localization accuracies. The last two measure are computed on the generated sentences, *i.e.*, given a test video segment, we perform the standard language generation inference and compute attention localization accuracy (no grounding measurement here because it is usually evaluated on GT sentences). We define $F1_{all}$ and $F1_{loc}$ as follows.

We define the number of object words in the generated sentences as A, the number of object words in the GT sentences as B, the number of correctly predicted object words in the generated sentences as C and the counterpart in the GT sentences as D, and the number of correctly predicted and localized words as E. A region prediction is considered correct if the object word is correctly predicted and also correctly localized (*i.e.*, IoU with GT box > 0.5).

In $F1_{all}$, the precision and recall can be defined as:

$$\text{Precision}_{all} = \frac{E}{A}, \quad \text{Recall}_{all} = \frac{E}{B} \tag{1}$$

However, since having box annotation for every single object in the scene is unlikely, an incorrectly-predicted word might not necessarily be a hallucinated object. Hence, we also compute $F1_{loc}$, which only considers correctly-predicted object words, i.e., only measures the localization quality and ignores errors result from the language generation. The precision and recall for $F1_{loc}$ are defined as:

$$\text{Precision}_{loc} = \frac{E}{C}, \quad \text{Recall}_{loc} = \frac{E}{D} \tag{2}$$

If multiple instances of the same object exist in the target sentence, we only consider the first instance. The precision and recall for the two metrics are computed for each object class, but it is set to zero if an object class has never been predicted. Finally, we average the scores by dividing by the total number of object classes in a particular split (val or test).

During model training, we restrict the grounding region candidates within the target frame (w/ GT box), *i.e.*, only consider the N_f proposals on the frame f with the GT box.

	$F1_{all}$		$F1_{loc}$	
Method	Precision	Recall	Precision	Recall
Unsup. (w/o SelfAttn)	3.76	3.63	12.6	12.9
Unsup.	0.28	0.27	1.13	1.13
Sup. Attn.	6.71	6.73	22.6	22.8
Sup. Grd.	6.25	5.84	21.2	21.2
Sup. Cls.	0.40	0.32	1.39	1.47
Sup. Attn.+Grd.	7.07	6.54	23.0	23.0
Sup. Attn.+Cls.	7.29	6.94	24.0	24.1
Sup. Grd. +Cls.	4.94	4.64	17.7	17.6
Sup. Attn.+Grd.+Cls.	7.42	6.81	23.7	23.9

Table 1: Attention precision and recall on generated sentences on ANet-Entities val set. All values are in %.

	$F1_{all}$		$F1_{loc}$	
Method	Precision	Recall	Precision	Recall
Unsup. (w/o SelfAttn)	3.62	3.85	11.7	11.8
Sup. Attn.+Cls.	7.64	7.55	25.1	24.8

Table 2: Attention precision and recall on generated sentences on ANet-Entities test set. All values are in %.

Method	B@1	B@4	M	C	S
Region Attn.	23.2	2.55	10.9	43.5	14.5
Tempo. Attn.	23.5	2.45	11.0	44.3	14.0
Both	**23.9**	**2.59**	**11.2**	**47.5**	**15.1**

Table 3: Ablation study for two attention modules using our best model. Results reported on val set.

A.3. Results on ActivityNet-Entities

We first include here the precision and recall associated with $F1_{all}$ and $F1_{loc}$ (see Tabs. 1, 2).

Temporal attention & region attention. We conduct ablation studies on the two attention modules to study the impact of each component on the overall performance (see Tab. 3). Each module alone performs similarly and the combination of two performs the best, which indicates the two attention modules are complementary. We hypothesize that the temporal attention captures the coarse-level details while the region attention captures more fine-grained details. Note that the region attention module takes in a lower sampling rate input than the temporal attention module, so

Method	λ_α	λ_β	λ_c	B@1	B@4	M	C	S	Attn.	Grd.	F1$_{all}$	F1$_{loc}$	Cls.
Unsup. (w/o SelfAttn)	0	0	0	70.0	27.5	22.0	60.4	15.9	22.0	25.9	4.44	12.8	17.6
Unsup.	0	0	0	69.3	26.8	22.1	59.4	15.7	4.04	16.3	0.80	2.09	1.35
Sup. Attn.	0.1	0	0	**71.0**	**28.2**	**22.7**	63.0	**16.3**	**42.3**	44.1	8.08	22.4	6.59
Sup. Grd.	0	0.1	0	70.1	27.6	22.5	**63.1**	16.1	38.5	49.5	7.59	21.0	0.03
Sup. Cls. (w/o SelfAttn)	0	0	1	70.1	27.6	22.0	60.2	15.8	20.9	32.1	4.12	11.5	**19.9**
Sup. Attn.+Grd.	0.1	0.1	0	70.2	27.6	22.5	62.3	**16.3**	**42.7**	49.8	**8.62**	**23.6**	0
Sup. Attn.+Cls.	0.1	0	1	70.0	27.9	22.6	62.4	**16.3**	42.1	46.5	**8.35**	**23.2**	**19.9**
Sup. Grd. +Cls.	0	0.1	1	70.4	28.0	**22.7**	62.8	**16.3**	29.0	**51.2**	5.19	13.7	19.7
Sup. Attn.+Grd.+Cls.	0.1	0.1	1	**70.6**	**28.1**	22.6	**63.3**	**16.3**	41.2	**50.8**	8.30	**23.2**	19.6

Table 4: Results on Flickr30k Entities val set. The top two scores on each metric are in bold.

	F1$_{all}$		F1$_{loc}$	
Method	Precision	Recall	Precision	Recall
Unsup. (w/o SelfAttn)	4.08	4.89	12.8	12.8
Unsup.	0.75	0.87	2.08	2.10
Sup. Attn.	7.46	8.83	22.4	22.5
Sup. Grd.	6.90	8.43	21.0	21.0
Sup. Cls. (w/o SelfAttn)	3.70	4.66	11.4	11.5
Sup. Attn.+Grd.	7.93	9.45	23.7	23.6
Sup. Attn.+Cls.	7.61	9.25	23.2	23.1
Sup. Grd. +Cls.	4.70	5.83	13.7	13.7
Sup. Attn.+Grd.+Cls.	7.56	9.20	23.2	23.2

Table 5: Attention precision and recall on generated sentences on Flickr30k Entities val set. All values are in %.

	F1$_{all}$		F1$_{loc}$	
Method	Precision	Recall	Precision	Recall
BUTD [1]	4.07	5.13	13.1	13.0
Our Unsup. (w/o SelfAttn)	3.44	4.47	11.6	11.8
Our Sup. Attn.+Grd.+Cls.	6.91	8.33	22.2	22.2

Table 6: Attention precision and recall on generated sentences on Flickr30k Entities test set. All values are in %.

we expect it can be further improved if having a higher sampling rate and the context (other events in the video). We leave this for future studies.

Notes on Video Paragraph Description. The authors of the SoTA method [7] kindly provided us with their result file and evaluation script, but as they were unable to provide us with their splits, we evaluated both methods on *our* test split. Even though we are under an unfair disadvantage, *i.e.*, the authors' val split might contain videos from our test split, we still outperform SotA method by a large margin, with relative improvements of 8.9-10% on all the metrics (as shown in Tab. 5).

Qualitative examples. See Figs. 3 and 4 for qualitative results of our methods and the Masked Transformer on ANet-Entities val set. We visualize the proposal with the highest attention weight in the corresponding frame. In (a), the supervised model correctly attends to "man" and "Christmas tree" in the video when generating the corresponding words. The unsupervised model mistakenly predicts "Two boys". In (b), both "man" and "woman" are correctly grounded. In (c), both "man" and "saxophone" are correctly grounded by our supervised model while Masked Transformer hallucinates a "bed". In (d), all the object words (*i.e.*, "people", "beach", "horses") are correctly localized. The caption generated by Masked Transformer is incomplete. In (e), surprisingly, not only major objects "woman" and "court" are localized, but also the small object "ball" is attended with a high precision. Masked Transformer incorrectly predicts the gender of the person. In (f), the Masked Transformer outputs an unnatural caption "A group of people are in a raft and a man in red raft raft raft raft raft" containing consecutive repeated words "raft".

A.4. Results on Flickr30k Entities

See Tab. 4 for the results on Flickr30k Entities val set. Note that the results on the test set can be found in the main paper in Tab. 4. The proposal upper bound for attention and grounding is 90.0%. For supervised methods, we perform a light hyper-parameter search and notice the setting $\lambda_\alpha = 0.1$, $\lambda_\beta = 0.1$ and $\lambda_c = 1$ generally works well. The supervised methods outperform the unsupervised baseline by a decent amount in all the metrics with only one exceptions: Sup. Cls., which has a slightly inferior result in CIDEr. The best supervised method outperforms the best unsupervised baseline by a relative 0.9-4.8% over all the metrics. The precision and recall associated with F1$_{all}$ and F1$_{loc}$ are shown in Tabs. 5, 6.

Qualitative examples. See Fig. 5 for the qualitative results by our methods and the BUTD on Flickr30k Entities val set. We visualize the proposal with the highest attention weight as the green box. The corresponding attention weight and the most confident object prediction of the proposal are displayed as the blue text inside the green box. In (a), the supervised model correctly attends to "man", "dog" and "snow" in the image when generating the corresponding words. The unsupervised model misses the word "snow" and BUTD misses the word "man". In (b), the supervised model successfully incorporates the detected visual clues (*i.e.*, "women", "building") into the description.

We also show a negative example in (c), where interestingly, the back of the chair looks like a laptop, which confuses our grounding module. The supervised model hallucinates a "laptop" in the scene.

A.5. Implementation Details

Region proposal and feature. We uniformly sample 10 frames per video segment (an event in ANet-Entities) and extract region features. For each frame, we use a Faster RCNN model [5] with a ResNeXt-101 FPN backbone [6] for region proposal and feature extraction. The Faster RCNN model is pretrained on the Visual Genonme dataset [3]. We use the same train-val-test split pre-processed by Anderson *et al*. [1] for joint object detection (1600 classes) and attribute classification. In order for a proposal to be considered valid, its confident score has to be greater than 0.2. And we limit the number of regions per image to a fixed 100 [2]. We take the output of the fc6 layer as the feature representation for each region, and fine-tune the fc7 layer and object classifiers with $0.1\times$ learning rate during model training.

Training details. We optimize the training with Adam (params: 0.9, 0.999). The learning rate is set to 5e-4 in general and to 5e-5 for fine-tuning, *i.e.*, fc7 layer and object classifiers, decayed by 0.8 every 3 epochs. The batch size is 240 for all the methods. We implement the model in PyTorch based on NBT[1] and train on 8x V100 GPUs. The training is limited to 40 epochs and the model with the best validation CIDEr score is selected for testing.

References

[1] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6077–6086, 2018. 3, 4

[2] Yu Jiang, Vivek Natarajan, Xinlei Chen, Marcus Rohrbach, Dhruv Batra, and Devi Parikh. Pythia v0. 1: the winning entry to the vqa challenge 2018. *arXiv preprint arXiv:1807.09956*, 2018. 4

[3] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73, 2017. 4

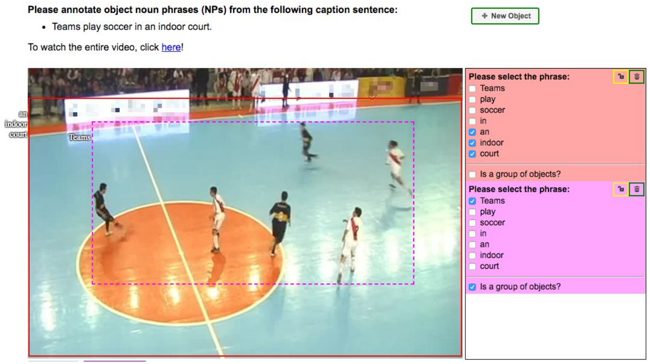
[4] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649, 2015. 1

[5] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015. 4

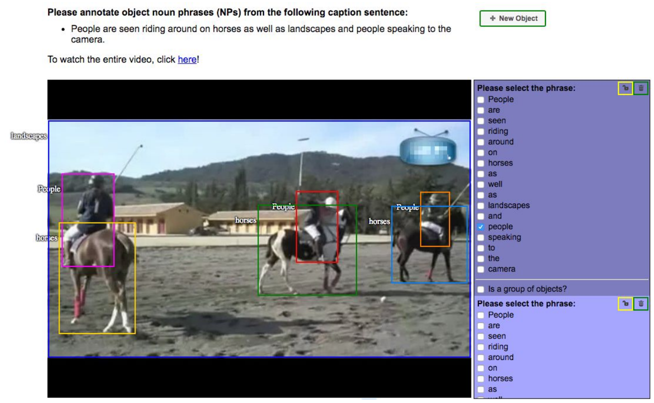
[6] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, pages 5987–5995. IEEE, 2017. 4

[7] Yilei Xiong, Bo Dai, and Dahua Lin. Move forward and tell: A progressive generator of video descriptions. *Proceedings of the European Conference on Computer Vision*, 2018. 3

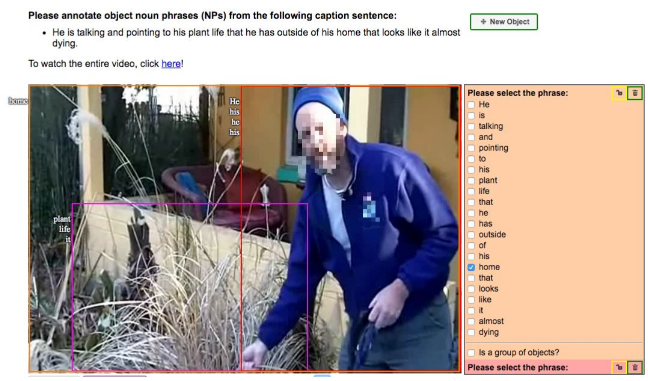
[1] https://github.com/jiasenlu/NeuralBabyTalk



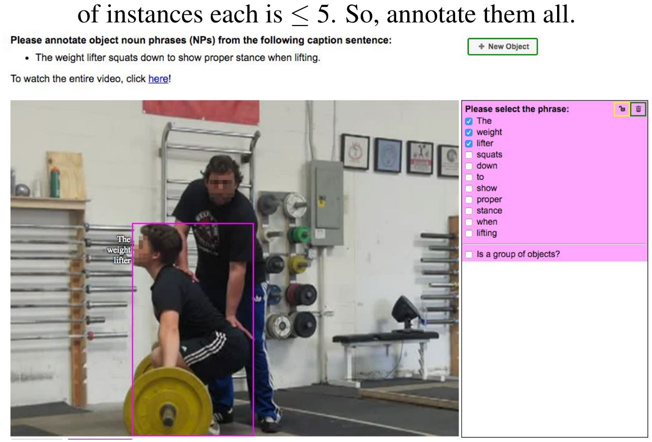
(a) "Teams" refers to more than 5 instances and hence should be annotated as a group.



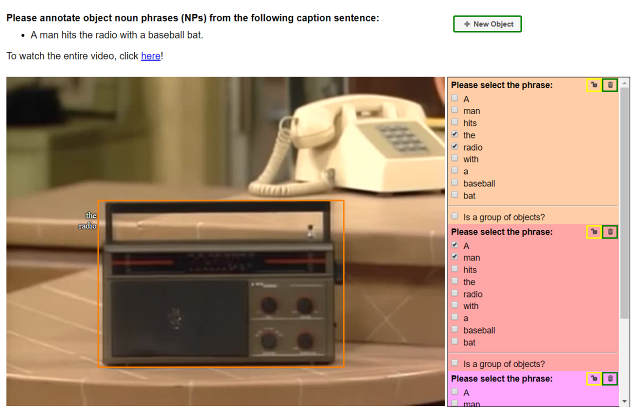
(b) "People" and "horses" can be clearly separated and the # of instances each is ≤ 5. So, annotate them all.



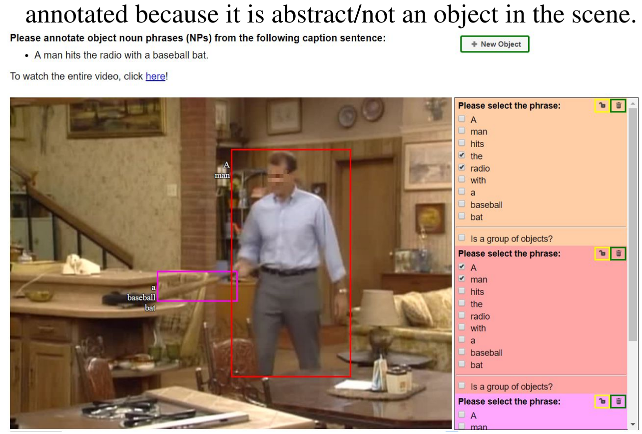
(c) "plant life" and "it" refer to the same box and "He", "'his", "he", "his" all refer to the same box.



(d) Only annotate the NP mentioned in the sentence, in this case, "The weight lifter". "proper stance" is a NP but not annotated because it is abstract/not an object in the scene.



(e) Note that (e) and (f) refer to the same video segment. See the caption of (f) for more details.



(f) "The radio" is annotated in a different frame as "a man" and "a baseball bat", since it cannot be clearly observed in the same frame.

Figure 1: Examples of our ActivityNet-Entities annotations in the annotation interface.

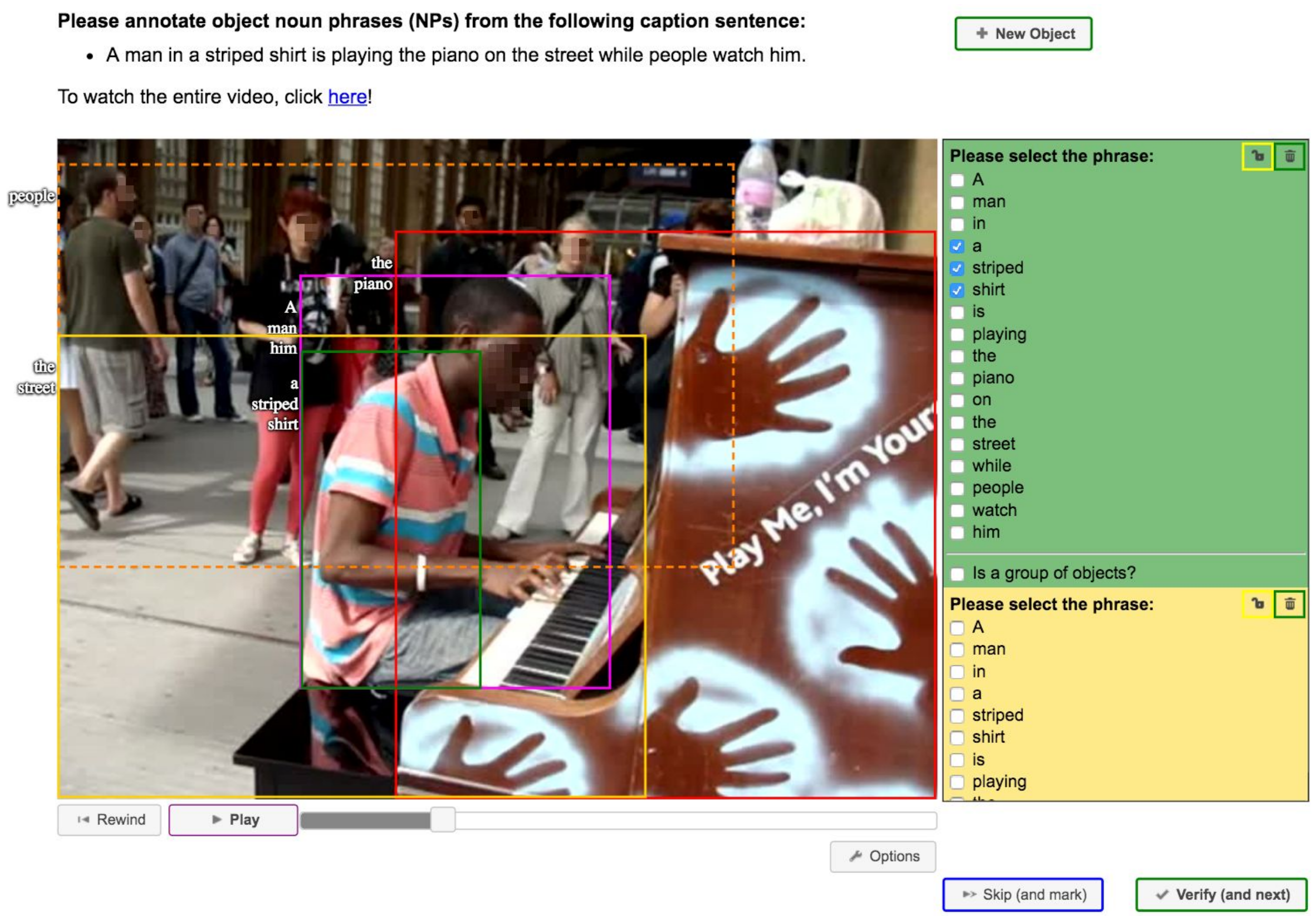


Figure 2: A screen shot of our annotation interface. The "verify (and next)" button indicates the annotation is under the verification mode, where the initial annotation is loaded and could be revised.

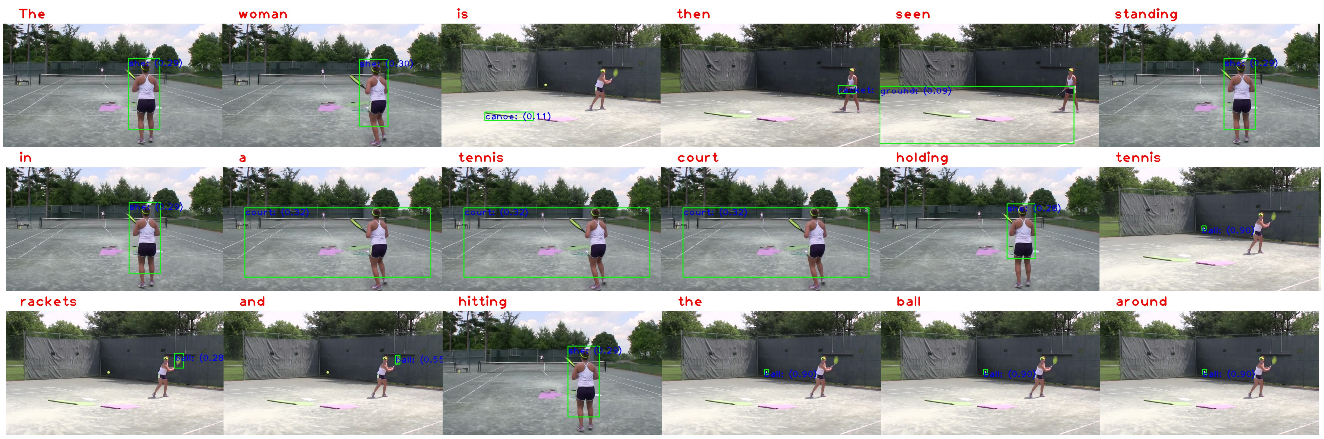
(a) **Sup.**: A man and a woman are standing in a room with a Christmas tree;
Unsup.: Two boys are seen standing around a room holding a tree and speaking to one another;
Masked Trans.: They are standing in front of the christmas tree;
GT: Then, a man and a woman set up a Christmas tree.

(b) **Sup.**: The man and woman talk to the camera;
Unsup.: The man in the blue shirt is talking to the camera;
Masked Trans.: The man continues speaking while the woman speaks to the camera;
GT: The man and woman continue speaking to the camera.

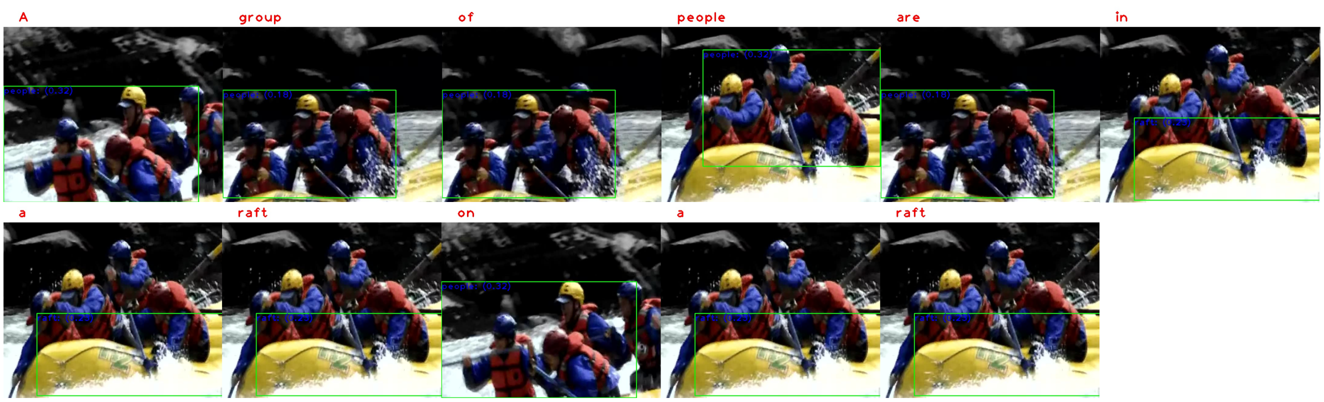
(c) **Sup.**: A man is standing in a room holding a saxophone;
Unsup.: A man is playing a saxophone;
Masked Trans.: A man is seated on a bed;
GT: We see a man playing a saxophone in front of microphones.

(d) **Sup.**: The people ride around the beach and ride around on the horses;
Unsup.: The people ride around the beach and ride around;
Masked Trans.: The camera pans around the area and the girl leading the horse and the woman leading the;
GT: We see four people on horses on the beach.

Figure 3: Qualitative results on ANet-Entities val set. The red text at each frame indicates the generated word. The green box indicates the proposal with the highest attention weight. The blue text inside the green box corresponds to i) the object class with the highest probability and ii) the attention weight. Better zoomed and viewed in color. See Sec. A.3 for discussion.

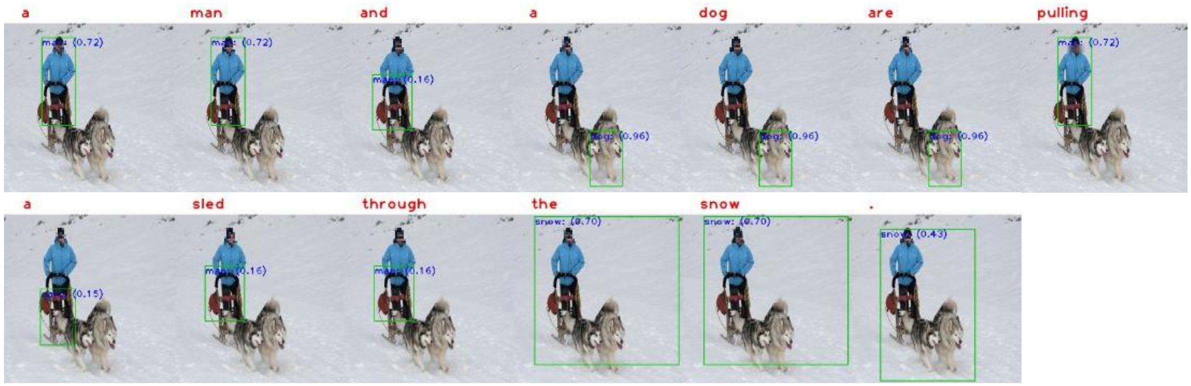


(e) **Sup.**: The woman is then seen standing in a tennis court holding tennis rackets and hitting the ball around;
Unsup.: The woman serves the ball with a tennis racket;
Masked Trans.: We see a man playing tennis in a court;
GT: Two women are on a tennis court, showing the technique to posing and hitting the ball.



(f) **Sup.**: A group of people are in a raft on a raft;
Unsup.: A group of people are in a raft;
Masked Trans.: A group of people are in a raft and a man in red raft raft raft raft raft;
GT: People are going down a river in a raft.

Figure 4: (Continued) Qualitative results on ANet-Entities val set. See the caption in Fig. 3 for more details.



(a) **Sup.**: A man and a dog are pulling a sled through the snow;
Unsup.: A man in a blue jacket is pulling a dog on a sled;
BUTD: Two dogs are playing in the snow;
GT (5): A bearded man wearing a blue jacket rides his snow sled pulled by his two dogs / Man in blue coat is being pulled in a dog sled by two dogs / A man in a blue coat is propelled on his sled by two dogs / A man us using his two dogs to sled across the snow / Two Huskies pull a sled with a man in a blue jacket.

(b) **Sup.**: Three women are standing in front of a building;
Unsup.: Three women in costumes are standing on a stage with a large wall in the background;
BUTD: Three women in yellow and white dresses are walking down a street;
GT (5): Three woman are crossing the street and on is wearing a yellow coat / Three ladies enjoying a stroll on a cold, foggy day / A woman in a yellow jacket following two other women / Three women in jackets walk across the street / Three women are crossing a street.

(c) **Sup.**: A man in a gray jacket is sitting in a chair with a laptop in the background;
Unsup.: A man in a brown jacket is sitting in a chair at a table;
BUTD: A man in a brown jacket is sitting in a chair with a woman in a brown jacket in a;
GT (5): Several chairs lined against a wall, with children sitting in them / A group of children sitting in chairs with monitors over them / Children are sitting in chairs under some television sets / Pre-teen students attend a computer class / Kids conversing and learning in class.

Figure 5: Qualitative results on Flickr30k Entities val set. Better zoomed and viewed in color. See Sec. A.4 for discussion.

__background__	egg	nail	kid	snowboard	hoop	roller	pasta
bagpipe	stilt	metal	butter	cheerleader	puck	kitchen	stage
coach	paper	dog	surfboard	landscape	scene	guitar	trophy
bull	dough	tooth	object	eye	scissors	grass	stone
rod	costume	pipe	ocean	sweater	ring	drum	swimmer
disc	oven	shop	person	camera	city	accordion	stand
dish	braid	shot	edge	vehicle	horse	ramp	road
chair	pinata	kite	bottle	raft	basketball	bridge	swimming
carpet	bunch	text	camel	themselves	monkey	wall	image
animal	group	barbell	photo	calf	top	soap	playground
gymnast	harmonica	biker	polish	teen	paint	pot	brush
mower	platform	shoe	cup	door	leash	pole	female
bike	window	ground	sky	plant	store	dancer	log
curler	soccer	tire	lake	glass	beard	table	area
ingredient	coffee	title	bench	flag	gear	boat	tennis
woman	someone	winner	color	adult	shorts	bathroom	lot
string	sword	bush	pile	baby	gym	teammate	suit
wave	food	wood	location	hole	wax	instrument	opponent
gun	material	tape	ski	circle	park	blower	head
item	number	hockey	skier	word	part	beer	himself
sand	band	piano	couple	room	herself	stadium	t-shirt
saxophone	they	goalie	dart	car	chef	board	cloth
team	foot	pumpkin	sumo	athlete	target	website	line
sidewalk	silver	hip	game	blade	instruction	arena	ear
razor	bread	plate	dryer	roof	tree	referee	he
clothes	name	cube	background	cat	bed	fire	hair
bicycle	slide	beam	vacuum	wrestler	friend	worker	slope
fence	arrow	hedge	judge	closing	iron	child	potato
sign	rock	bat	lady	male	coat	bmx	bucket
jump	side	bar	furniture	dress	scuba	instructor	cake
street	everyone	artist	shoulder	court	rag	tank	piece
video	weight	bag	towel	goal	clip	hat	pin
paddle	series	she	gift	clothing	runner	rope	intro
uniform	fish	river	javelin	machine	mountain	balance	home
supplies	gymnasium	view	glove	rubik	microphone	canoe	ax
net	logo	set	rider	tile	angle	it	face
exercise	girl	frame	audience	toddler	snow	surface	pit
body	living	individual	crowd	beach	couch	player	cream
trampoline	flower	parking	people	product	equipment	cone	lemon
leg	container	racket	back	sandwich	chest	violin	floor
surfer	house	close	sponge	mat	contact	helmet	fencing
water	hill	arm	mirror	tattoo	lip	shirt	field
studio	wallpaper	reporter	diving	ladder	tool	paw	other
sink	dirt	its	slice	bumper	spectator	bowl	oar
path	toy	score	leaf	end	track	member	picture
box	cookie	finger	bottom	baton	flute	belly	frisbee
boy	guy	teens	tube	man	cigarette	vegetable	lens
stair	card	pants	ice	tomato	mouth	pan	pool
bow	yard	opening	skateboarder	neck	letter	wheel	building
credit	skateboard	screen	christmas	liquid	darts	ball	lane
smoke	thing	outfit	knife	light	pair	drink	phone
trainer	swing	toothbrush	hose	counter	knee	hand	mask
shovel	castle	news	bowling	volleyball	class	fruit	jacket
kayak	cheese	tub	diver	truck	lawn	student	stick

Table 7: List of objects in ActivityNet-Entities, including the "__background__" class.