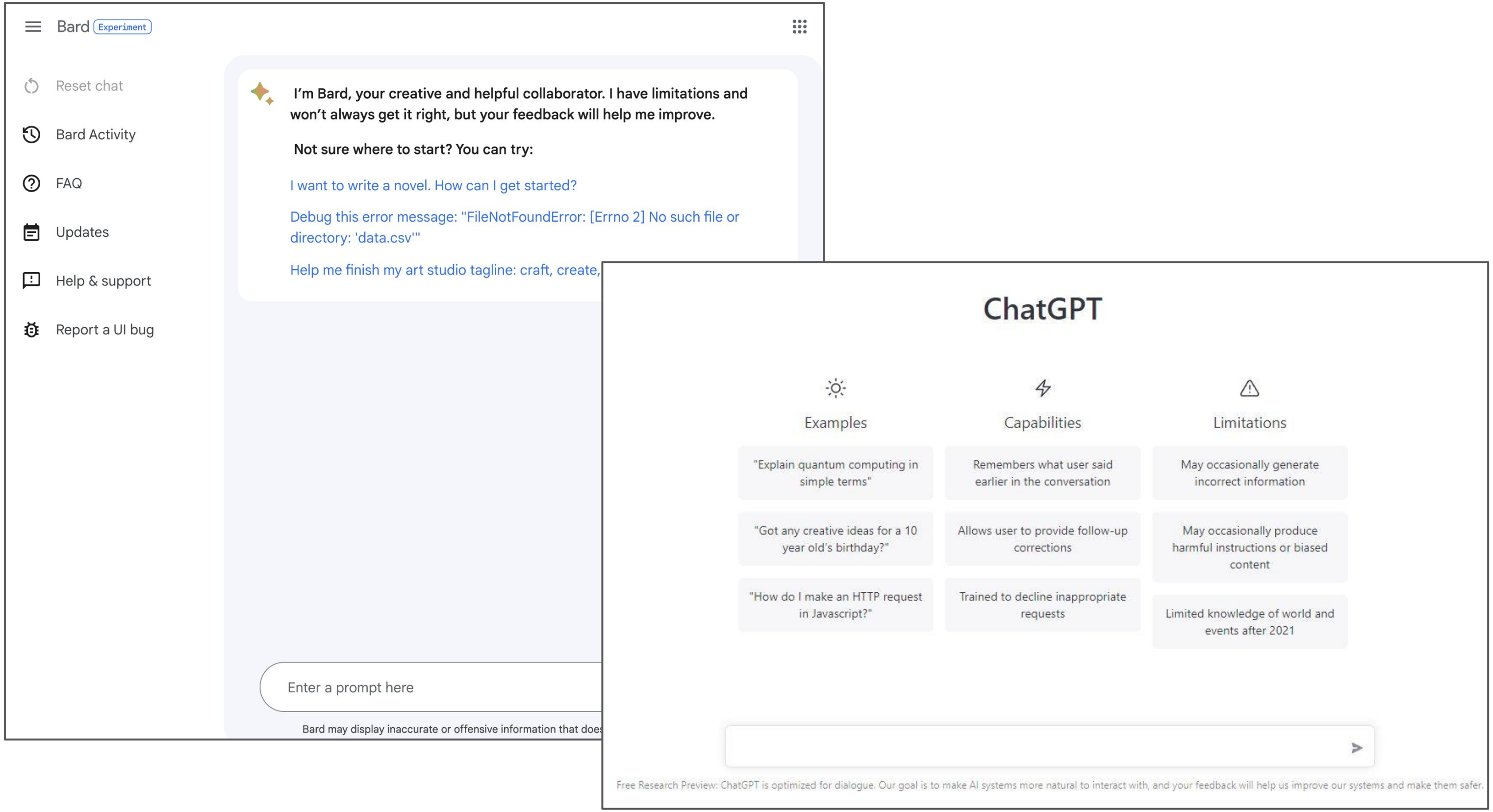
Dense Video Object Captioning from Disjoint Supervision

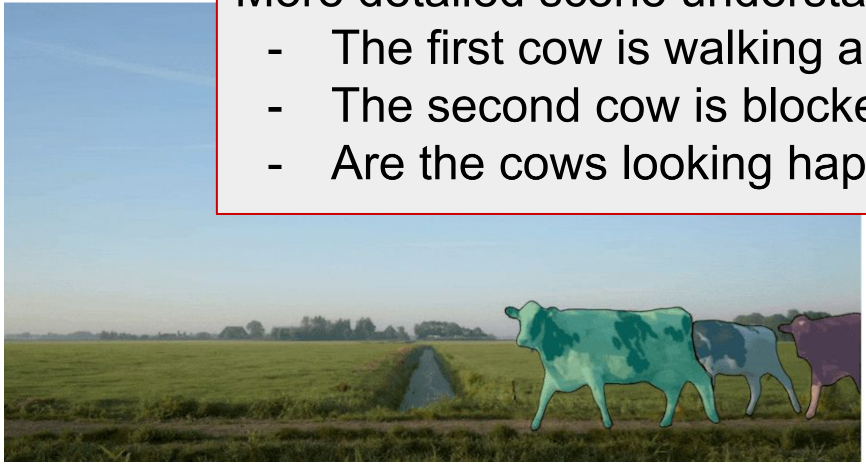
Xingyi Zhou*, Anurag Arnab*, Chen Sun, Cordelia Schmid

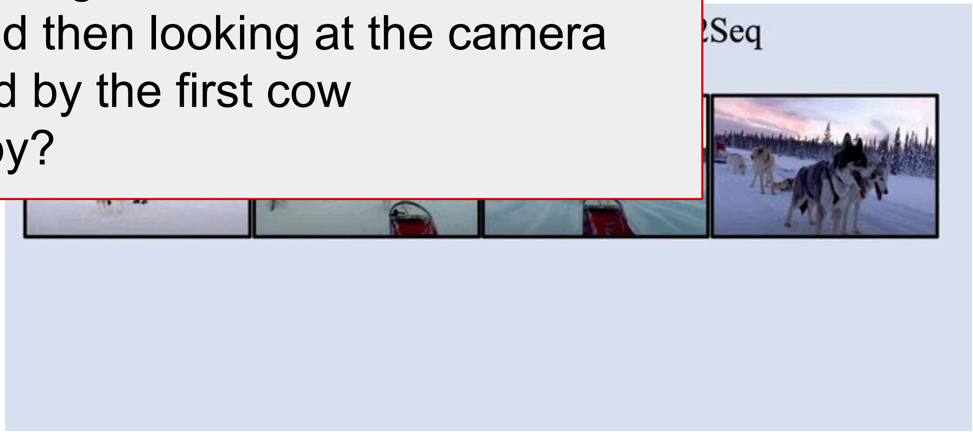
Recent advances in language



Recent advances in vision

More detailed scene understanding:
- The first cow is walking and then looking at the camera
- The second cow is blocked by the first cow
- Are the cows looking happy?





Segment any object

Describe the whole video

Video credit: https://ai.facebook.com/blog/segment-anything-foundation-model-image-segmentation/

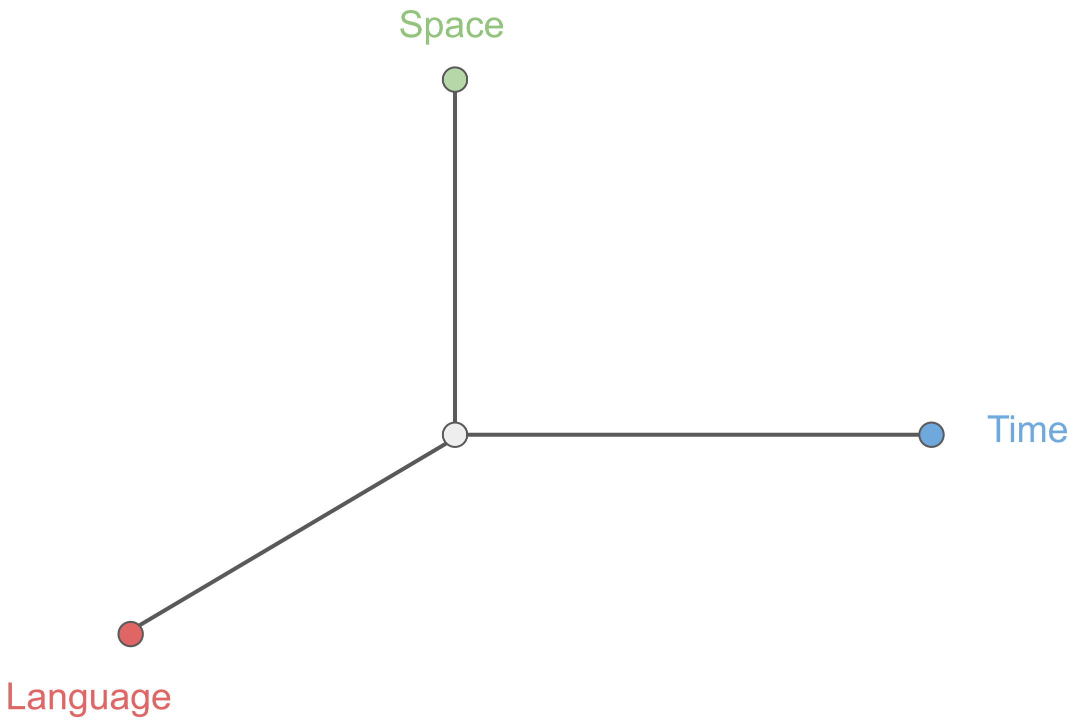
Video credit: https://ai.googleblog.com/2023/03/vid2seq-pretrained-visual-language.html

Goal: detect, track, and describe all objects in the video

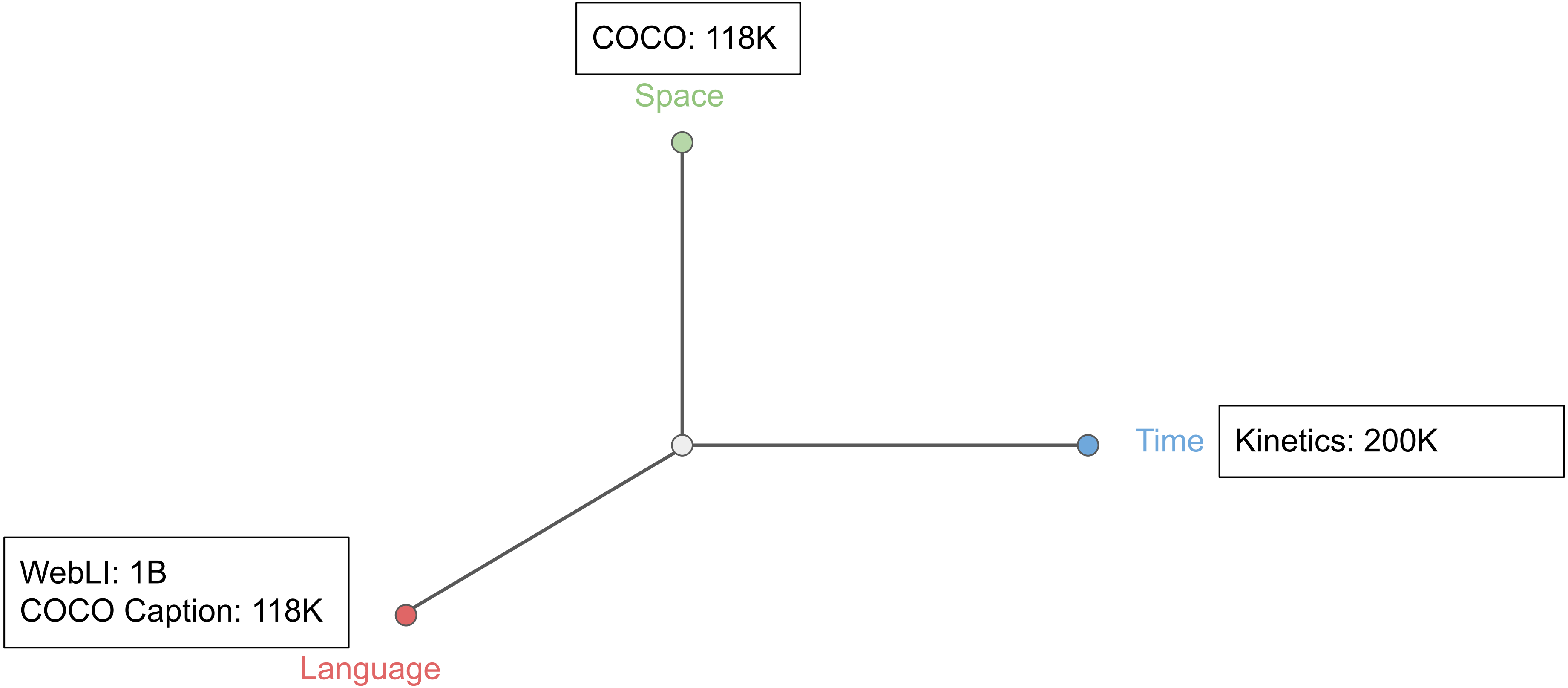
Applications:

- Detailed video description/ caption
- Video object grounding
- Video question answering

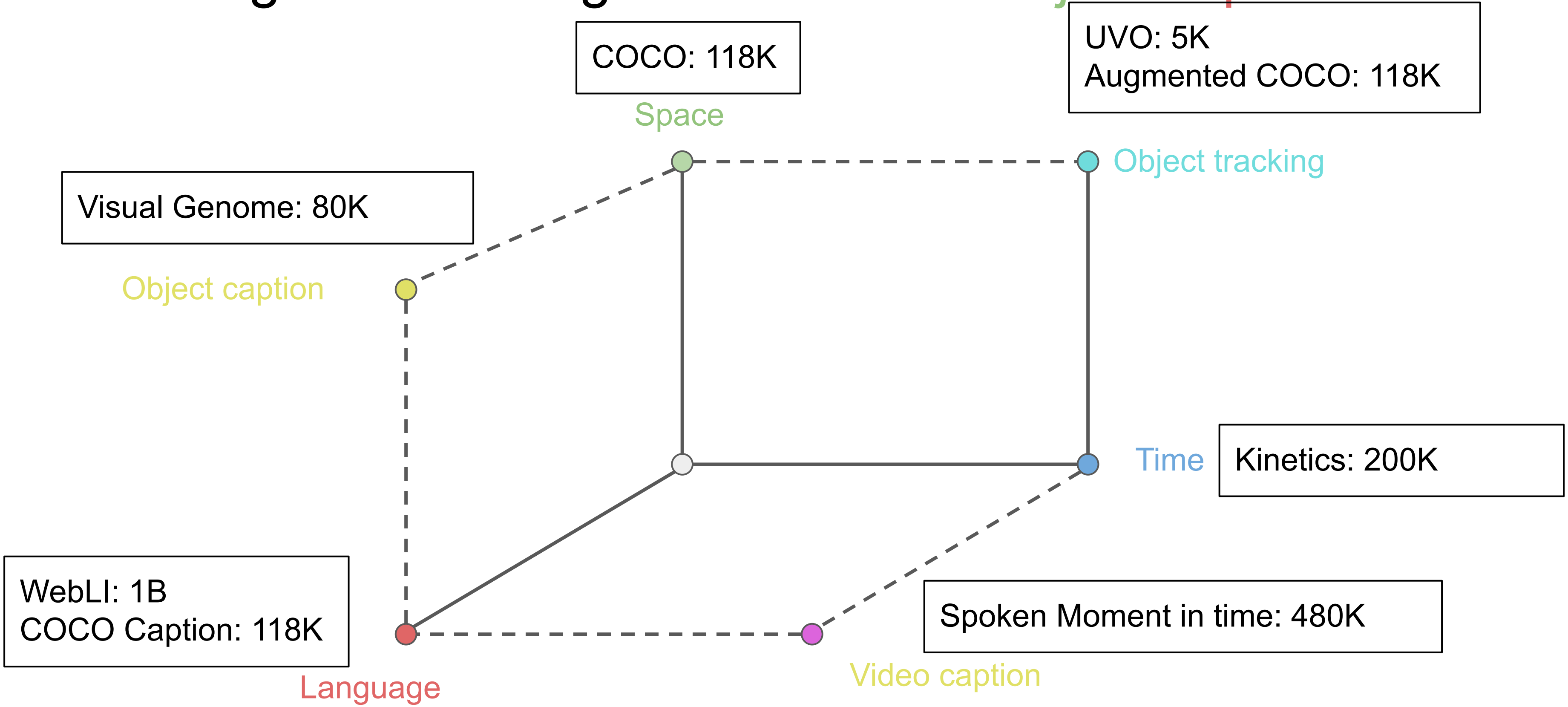
Challenge #1: training data: Tracked object captions



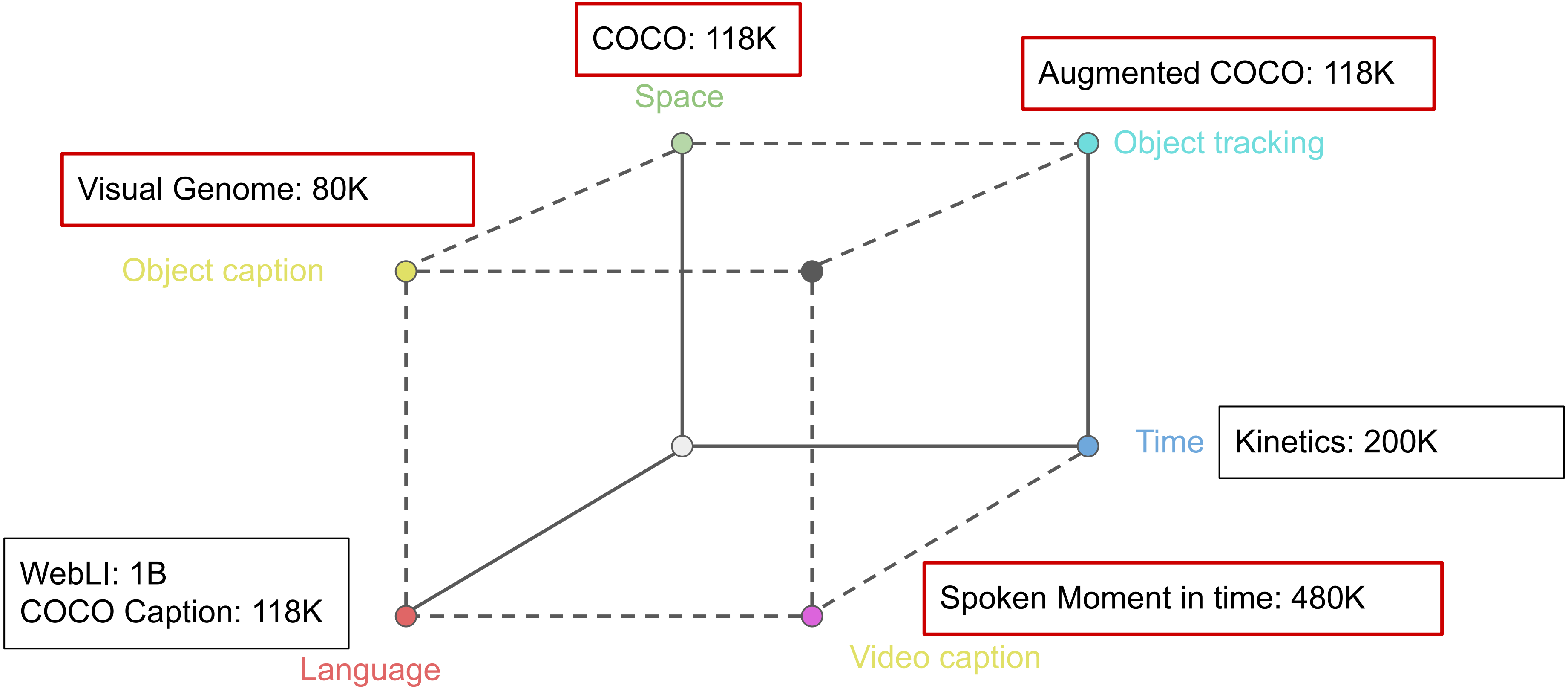
Challenge #1: training data: Tracked object captions

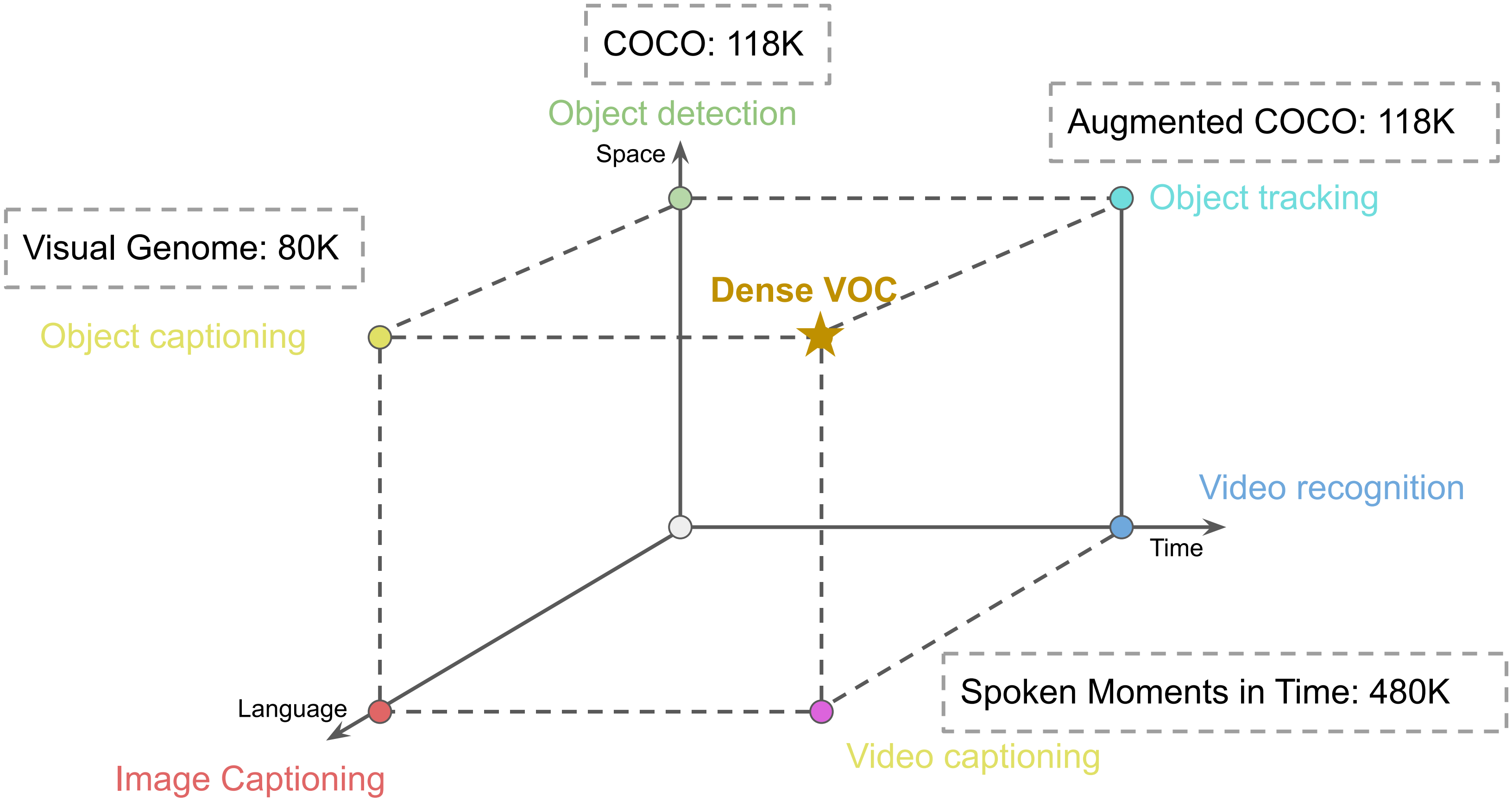


Challenge #1: training data: Tracked object captions



Our idea: disjoint weakly-supervised training

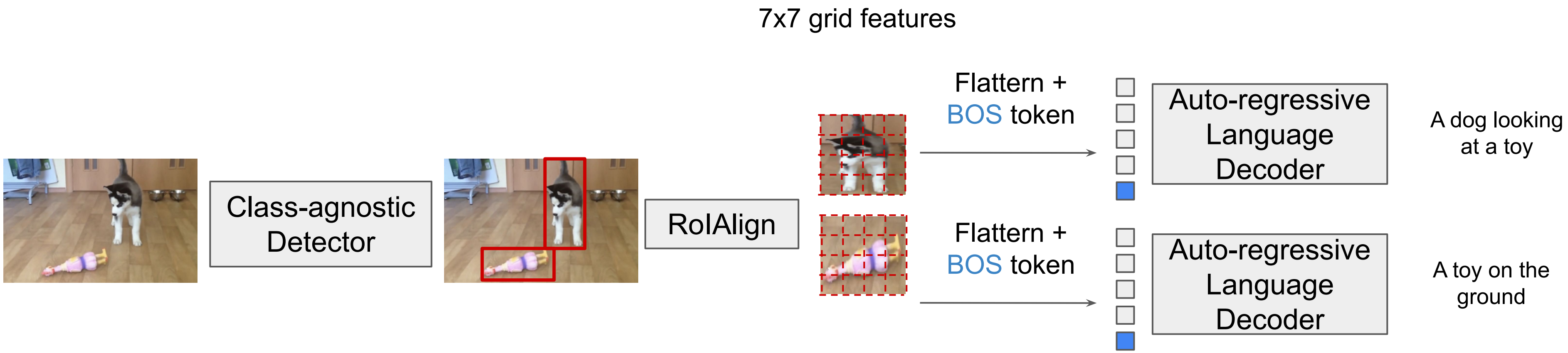



COCO: 118K
Object detection
Space
Augmented COCO: 118K
Object tracking
Visual Genome: 80K
Dense VOC
Object captioning
Video recognition
Time
Language
Spoken Moments in Time: 480K
Video captioning
Image Captioning

Challenge #2: model

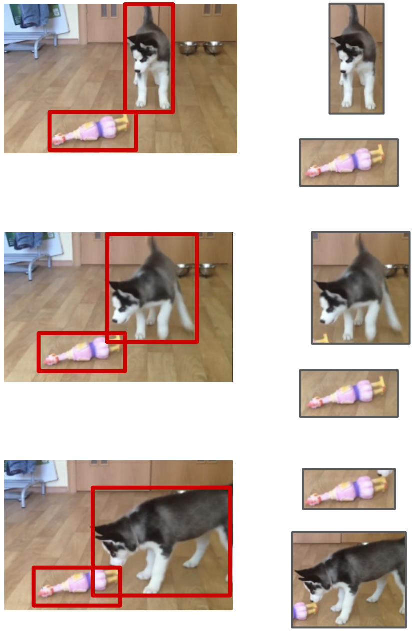
Dense object captions
→ **Tracked** dense object captions in video

Preliminary: dense object caption

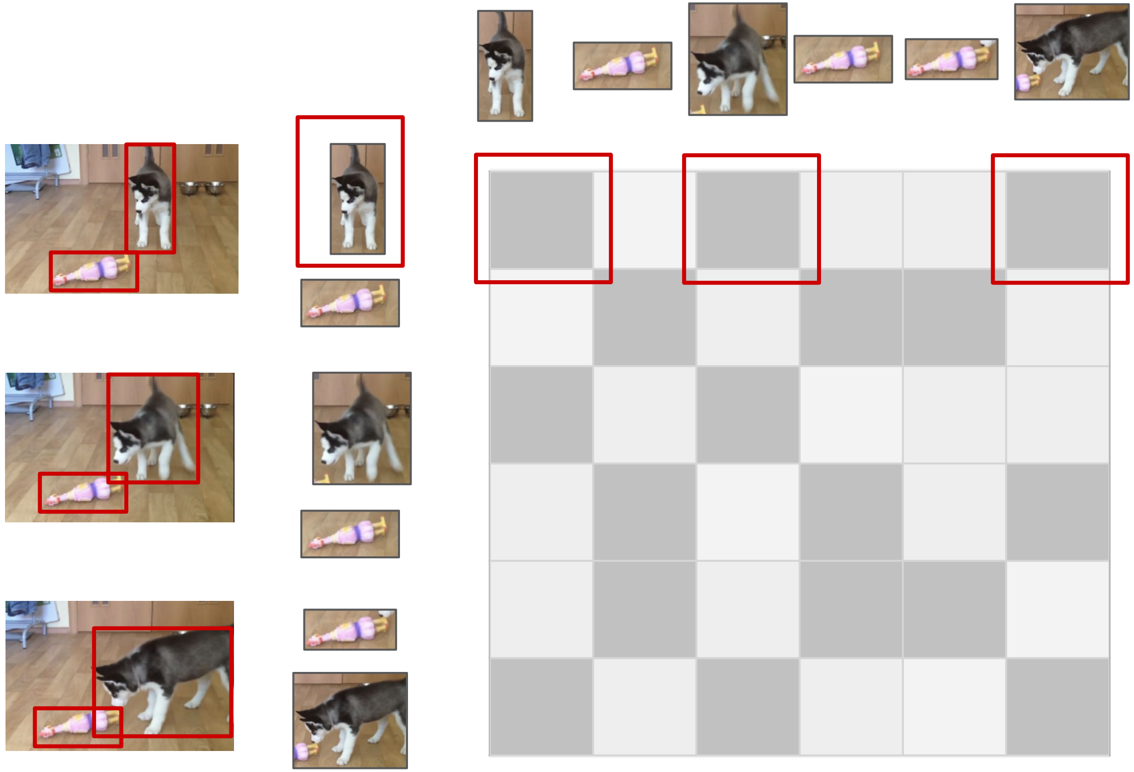


Wu et al, GRiT: A Generative Region-to-text Transformer for Object Understanding, arXiv 2022

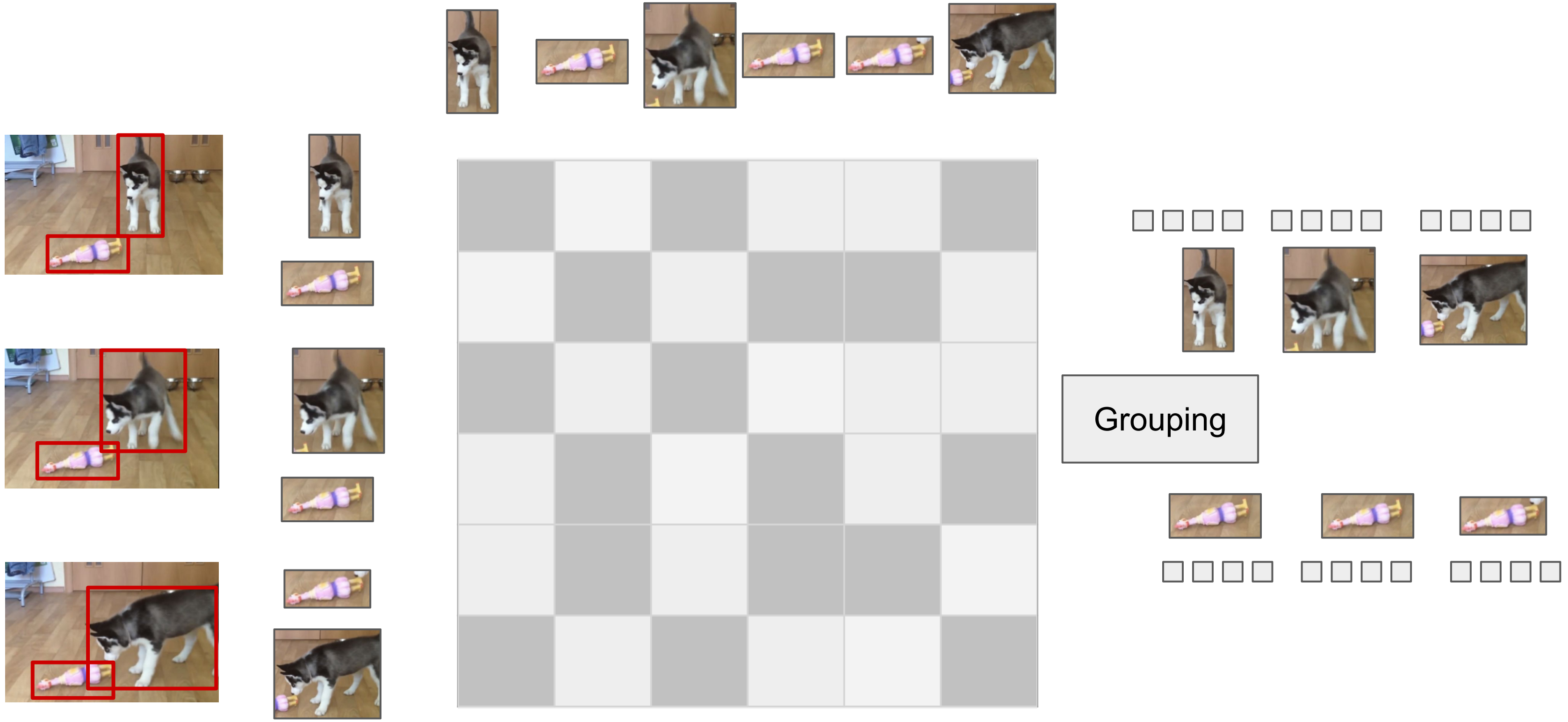
End-to-end video object tracking and caption



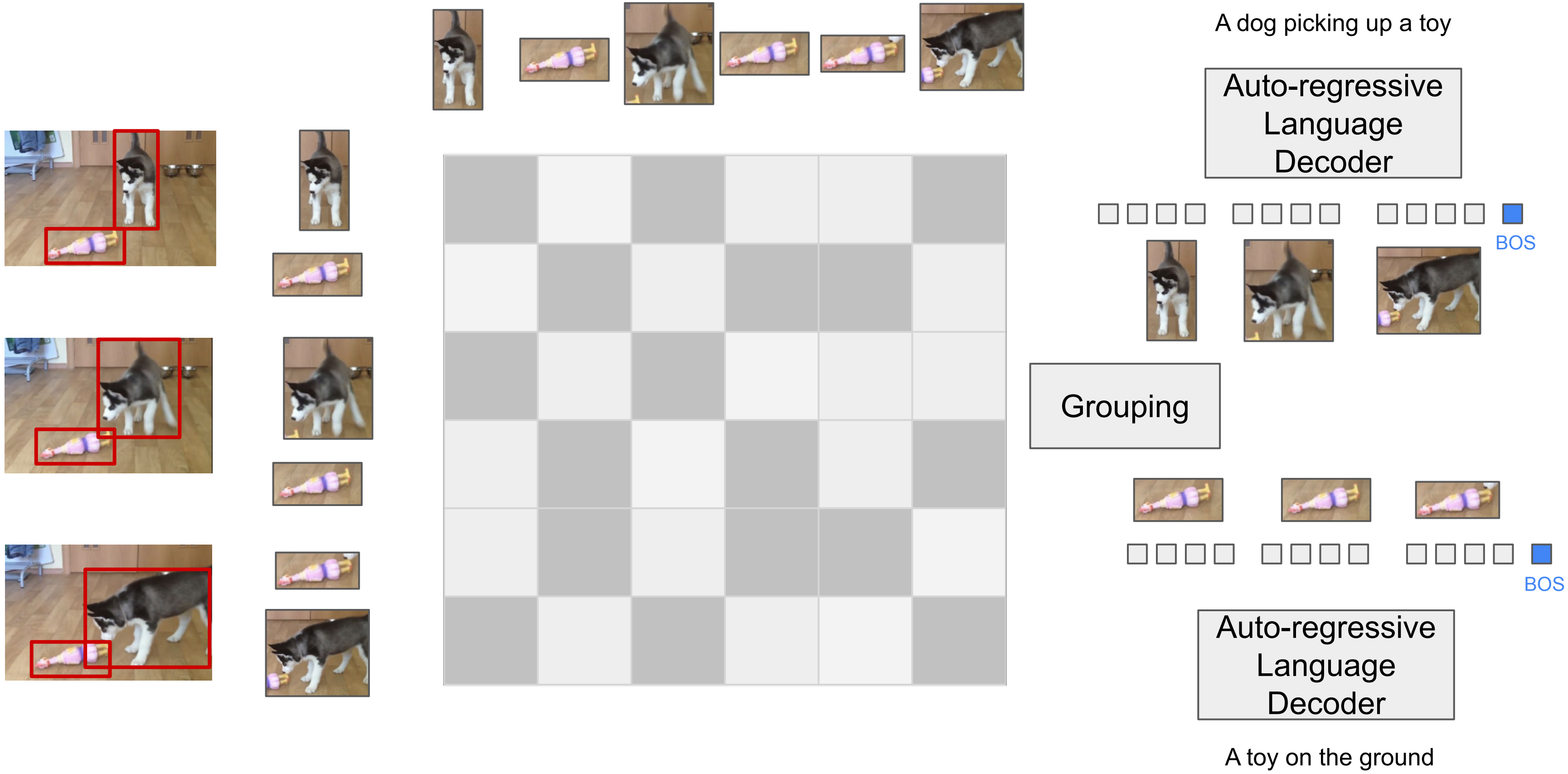
End-to-end video object tracking and caption



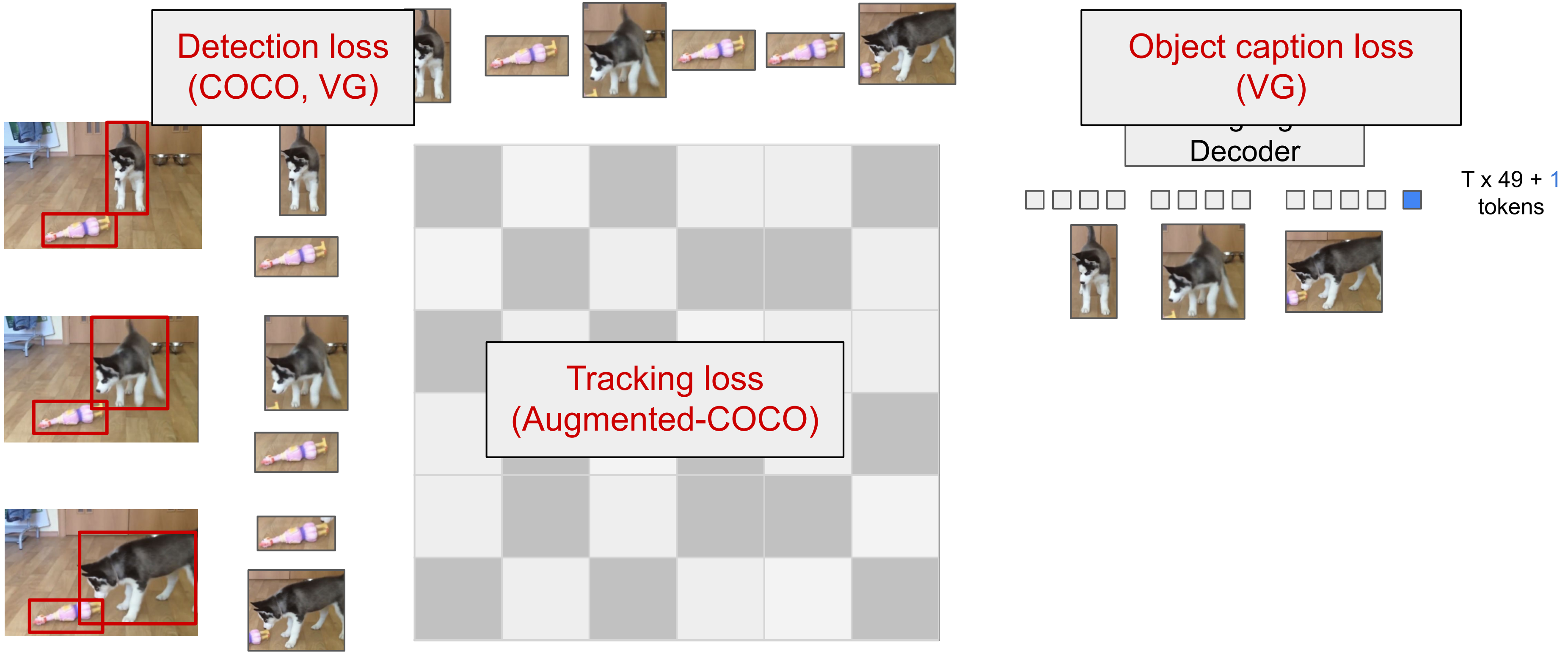
End-to-end video object tracking and caption



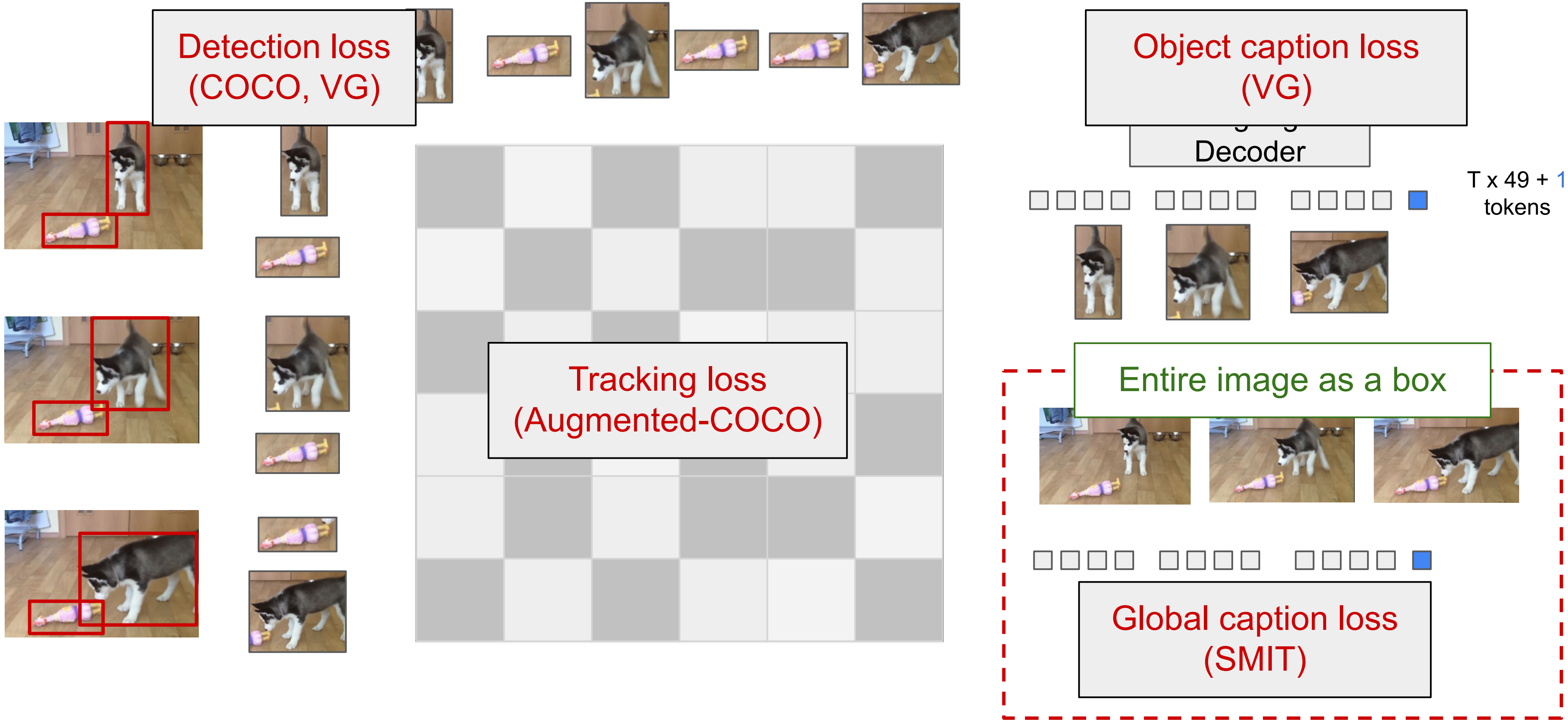
End-to-end video object tracking and caption



Training with disjoint annotation

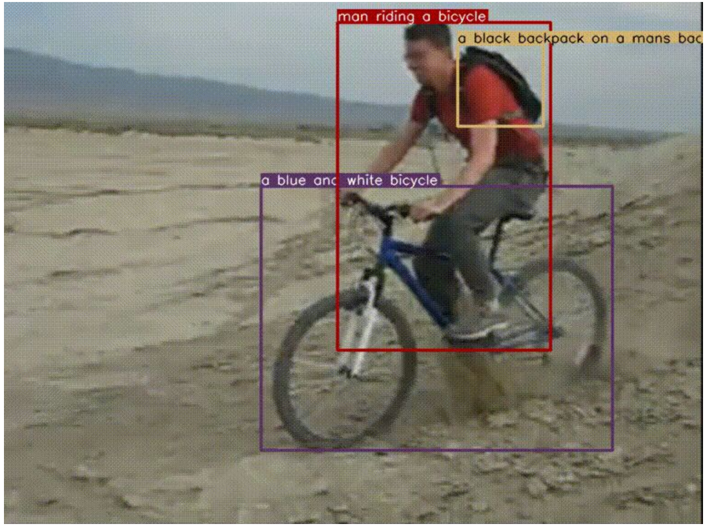


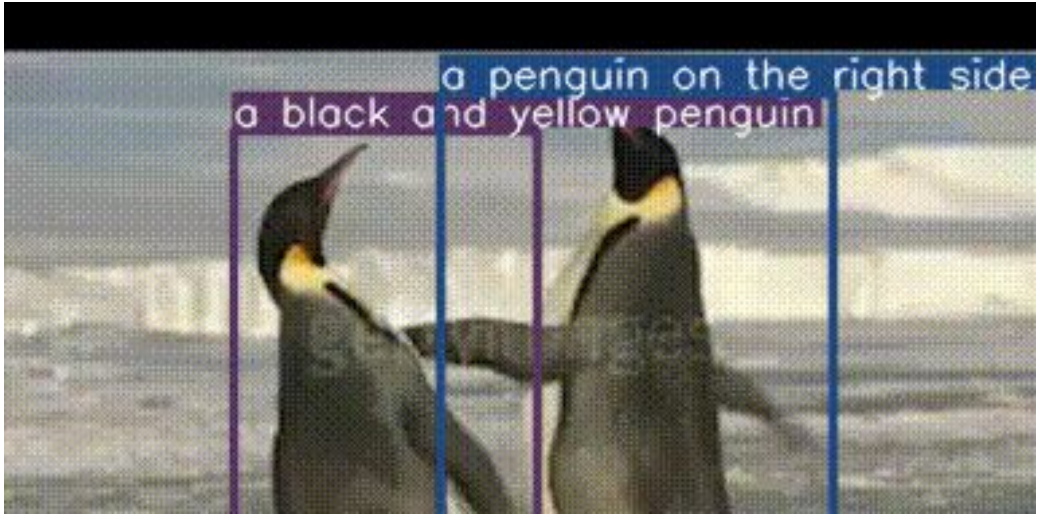
Training on video global caption data (weakly-supervised)



Results

Qualitative evaluation on SMiT

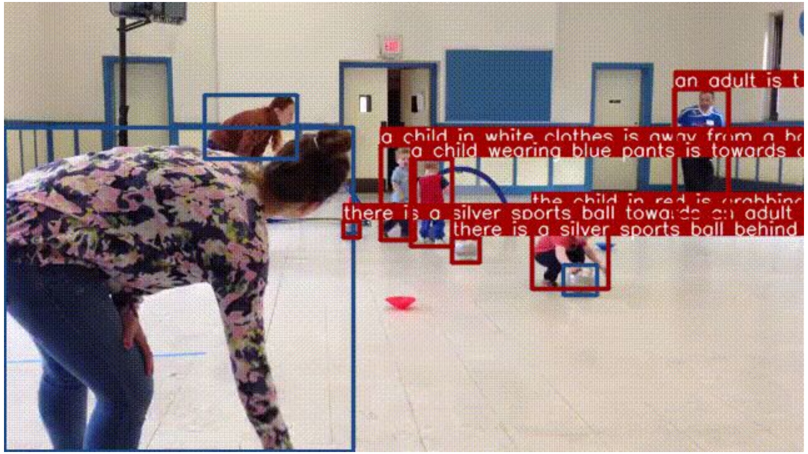




[More results]

Fine-tuning/ evaluation dataset - VidSTG

- Originally designed for spatial-temporal video object grounding.
- 5K training videos, 600 validation videos.
- We re-purpose the annotation for dense-video object caption.



VidSTG example annotation

Fine-tuning/ evaluation dataset - Video localized narrative

- Originally designed for spatial video object grounding.
- 5K/ 2K training/ validation videos.
- We re-purpose the annotation for dense-video object caption.

VLN example annotation

Disjoint multi-dataset training enables zero-shot application

#	COCO	VG	SMiT	Aug-COCO	VidSTG					VLN				
					CHOTA	DetA	AssA	CapA	AP_M	CHOTA	DetA	AssA	CapA	AP_M
1	✓				-	50.5	-	-	-	-	27.9	-	-	-
2		✓			-	18.9	-	9.4	18.7	-	12.2	-	7.2	10.1
3			✓		-	-	-	-	-	-	-	-	-	-
4		✓	✓		-	20.8	-	**10.2**	19.2	-	14.1	-	9.1	12.2
5	✓	✓			-	41.5	-	9.4	38.3	-	26.7	-	7.7	19.4
6	✓		✓		-	**51.7**	-	4.0	35.2	-	**28.7**	-	6.4	15.9
7	✓	✓	✓		-	45.1	-	9.8	39.0	-	24.4	-	8.2	19.9
8	✓	✓	✓	✓	**28.0**	45.9	**48.0**	10.0	**39.8**	**28.2**	28.6	**83.4**	**9.4**	**21.3**

Disjoint multi-dataset pre-training helps fine-tuning

+5.2

#	COCO	VG	SMiT	Aug-COCO	VidSTG CHOTA	VidSTG DetA	VidSTG AssA	VidSTG CapA	VidSTG AP_M	VLN CHOTA	VLN DetA	VLN AssA	VLN CapA	VLN AP_M
0					49.3	61.2	54.8	35.8	64.1	29.5	35.4	82.5	8.8	35.1
1	✓				51.6	65.6	56.9	36.8	69.3	31.4	43.9	86.2	8.2	36.3
2		✓			52.3	64.6	58.4	38.0	68.7	39.8	**45.1**	84.2	16.6	45.9
			✓		46.9	60.3	40.0	33.0	59.3	36.1	41.2	78.9	14.5	38.0
4		✓	✓		52.7	64.4	58.1	39.1	70.0	40.2	44.2	83.4	**17.6**	47.2
5	✓	✓			54.3	**66.6**	**61.0**	39.6	**71.2**	40.4	44.1	86.6	17.2	46.1
6	✓		✓		51.9	65.2	57.7	37.1	69.6	35.3	43.7	86.1	11.6	41.3
7	✓	✓	✓		**54.5**	66.1	60.2	**40.7**	**71.2**	**40.8**	44.2	**87.0**	**17.6**	**48.2**
8	✓	✓	✓	✓	54.3	66.1	59.8	40.5	**71.2**	**40.8**	44.3	86.9	**17.6**	**48.2**

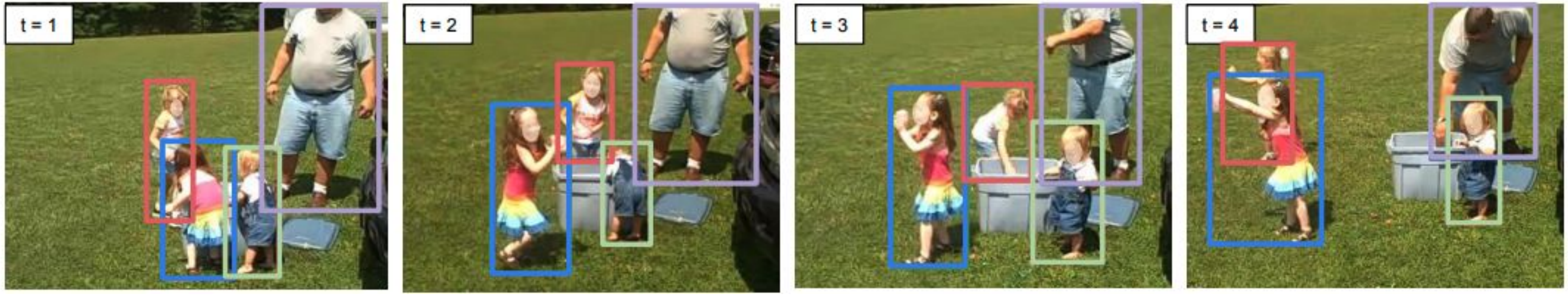
Importance of the tracking module

Per-frame caption
Caption Switch: 38.1%
Caption Accuracy: 37.1

Trajectory caption
Caption Switch: **17.5%**
Caption Accuracy: **38.0**

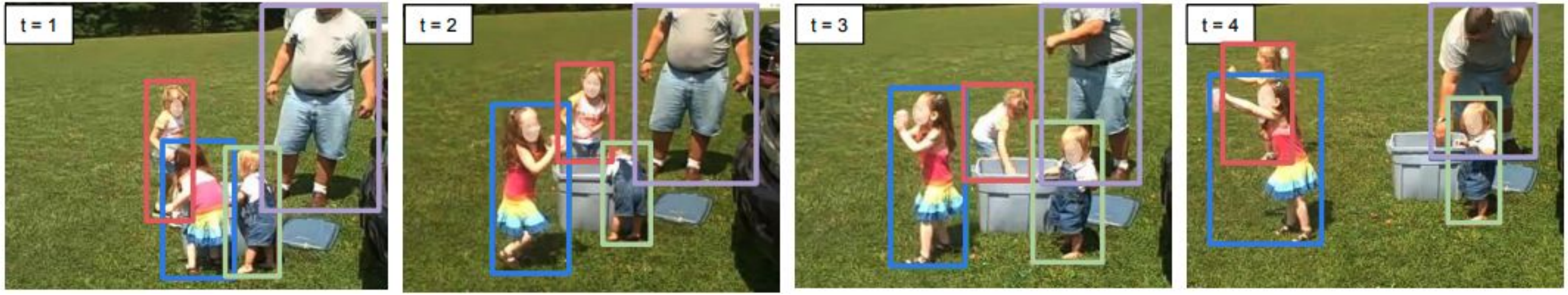
Application to video grounding

Query: q = “A child holds a toy on the grass”



Application to video grounding

Query: q = “A child holds a toy on the grass”

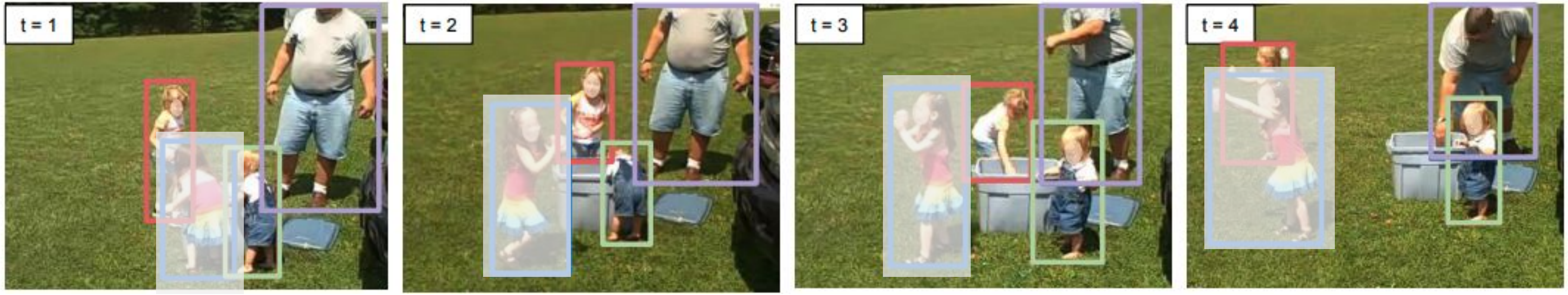


likelihood(, q) = 0.9 likelihood(, q) = 0.5

likelihood(, q) = 0.4 likelihood(, q) = 0.1

Application to video grounding

Query: q = “A child holds a toy on the grass”



likelihood(, q) = 0.9

likelihood(, q) = 0.5

likelihood(, q) = 0.4

likelihood(, q) = 0.1

Grounding results

	Recall	Precision	Both
ReferFormer [55, 61]	66.7	53.9	48.3
GRiT [60]	77.4	**68.5**	62.1
Ours	**86.7**	68.4	**65.1**

VLN spatial-grounding

	Finetuned	Zero-shot
STVGBert [52]	47.3	-
TubeDETR [66]	59.0	-
STCAT [29]	61.7	-
Ours	**61.9**	54.1

VidSTG spatial-grounding

Takeaway

- We propose a new task of dense video object captioning.
- We can train this task on large datasets with incomplete disjoint annotations.
- We propose an end-to-end tracking-and-caption framework that produces consistent captions.
- Our model can directly apply to video grounding tasks with state-of-the-art performance.