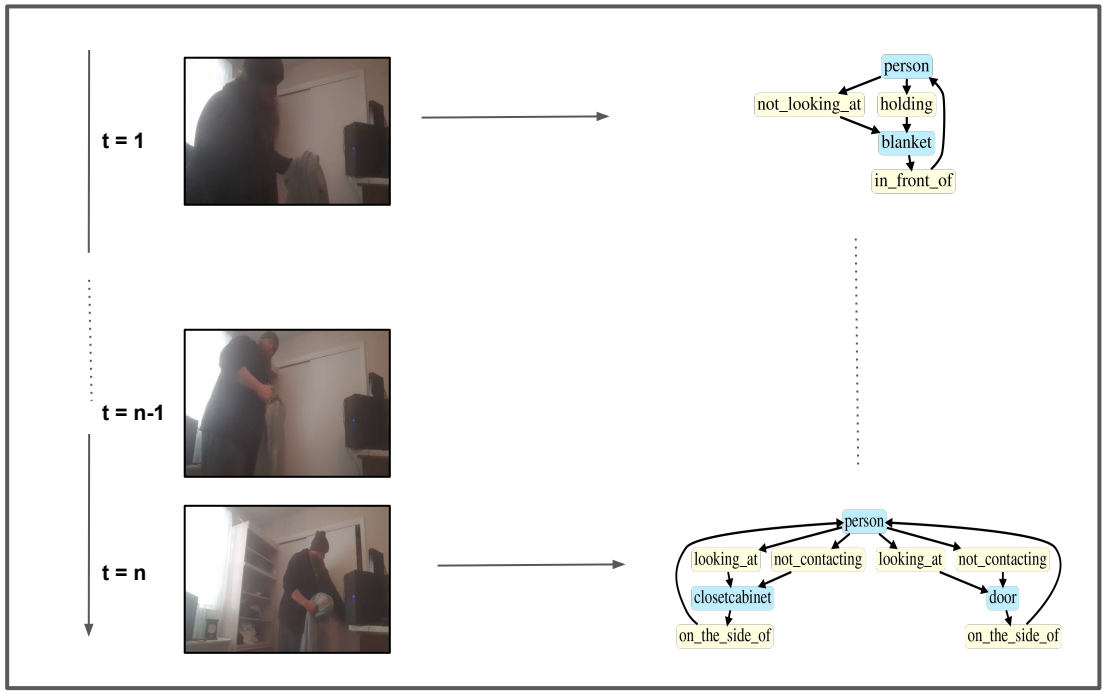
Scene Graph **and** Videos

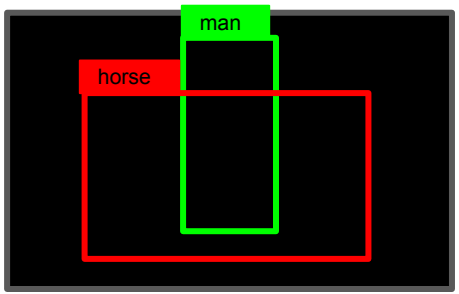


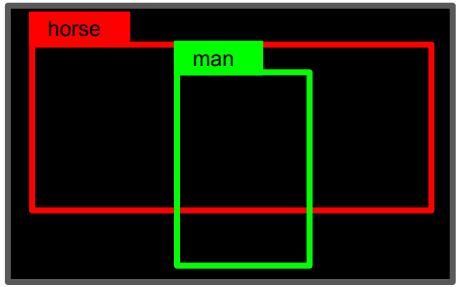
Contents

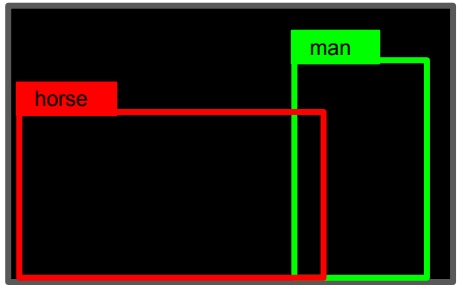
- What is ~~Seen~~ Scene Graph (SG) ? [5 mins]
- How to Generate SG ? [10 mins]
- Spatio-temporal SG [5 mins]
- Experiments with Spatio-remporal SG [5 mins]
- Plans for future + Questions [--]

What is SG ?

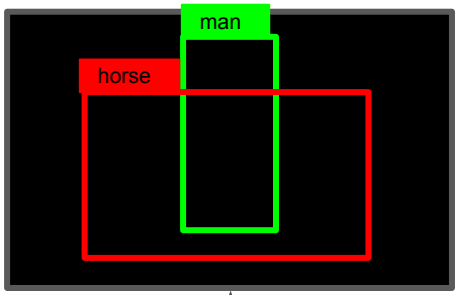
Puzzle



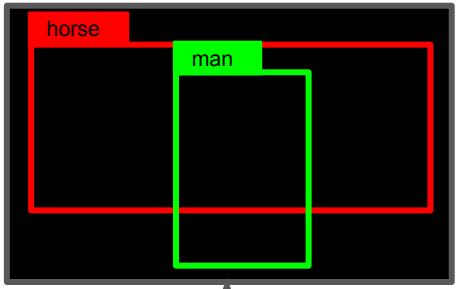




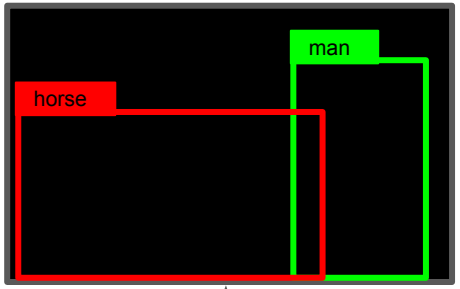
Puzzle



man **riding** horse

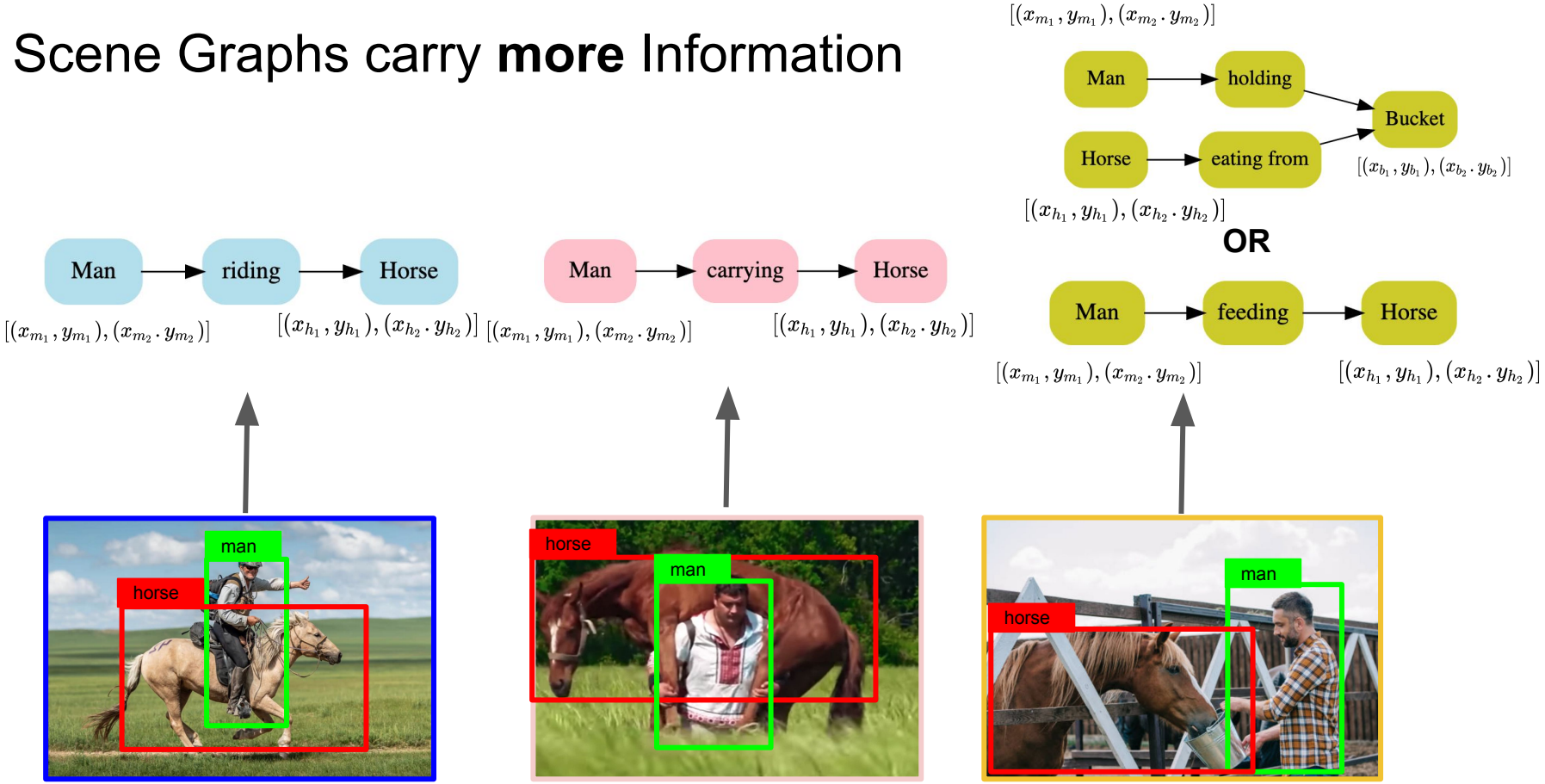


man **carrying** horse

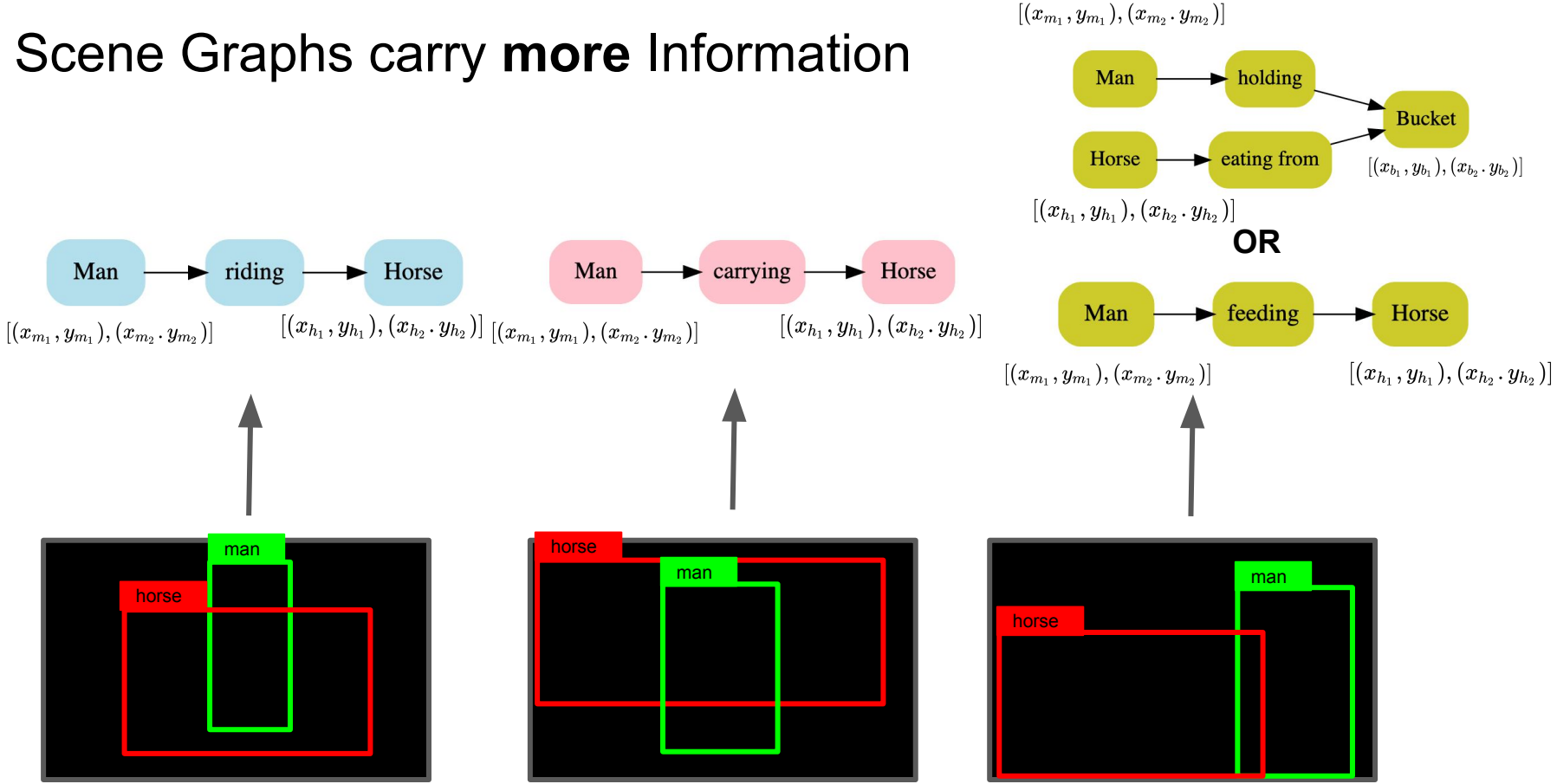


man **feeding** horse

Scene Graphs carry **more** Information



Scene Graphs carry **more** Information



Attributes and relationships help recognize objects

[1] Biederman, Visual Memory: What do you know about what you saw? Cognitive Psychology, 1982

Captioning the image

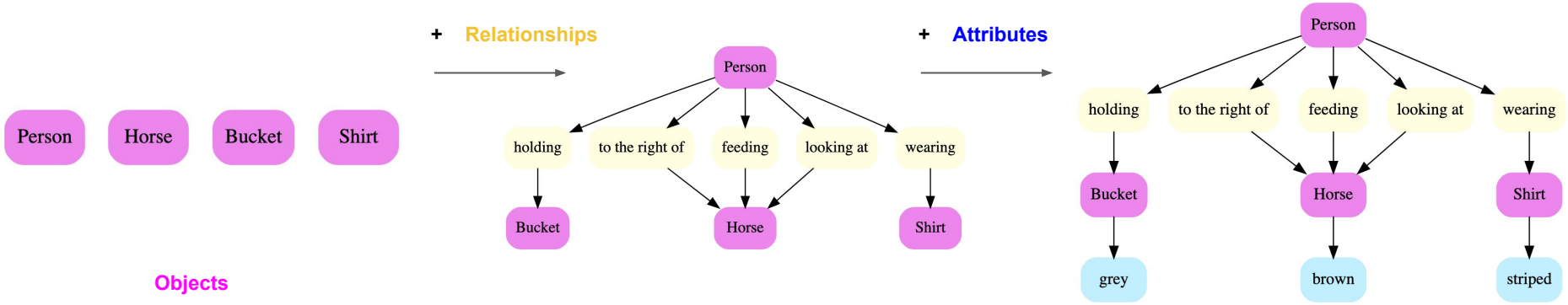
- How do people caption this image ?
  - A man is feeding the horse
  - Man in stripled shirt feeding the brown horse
  - A brown horse to the left of man is being fed through a bucket

Objects

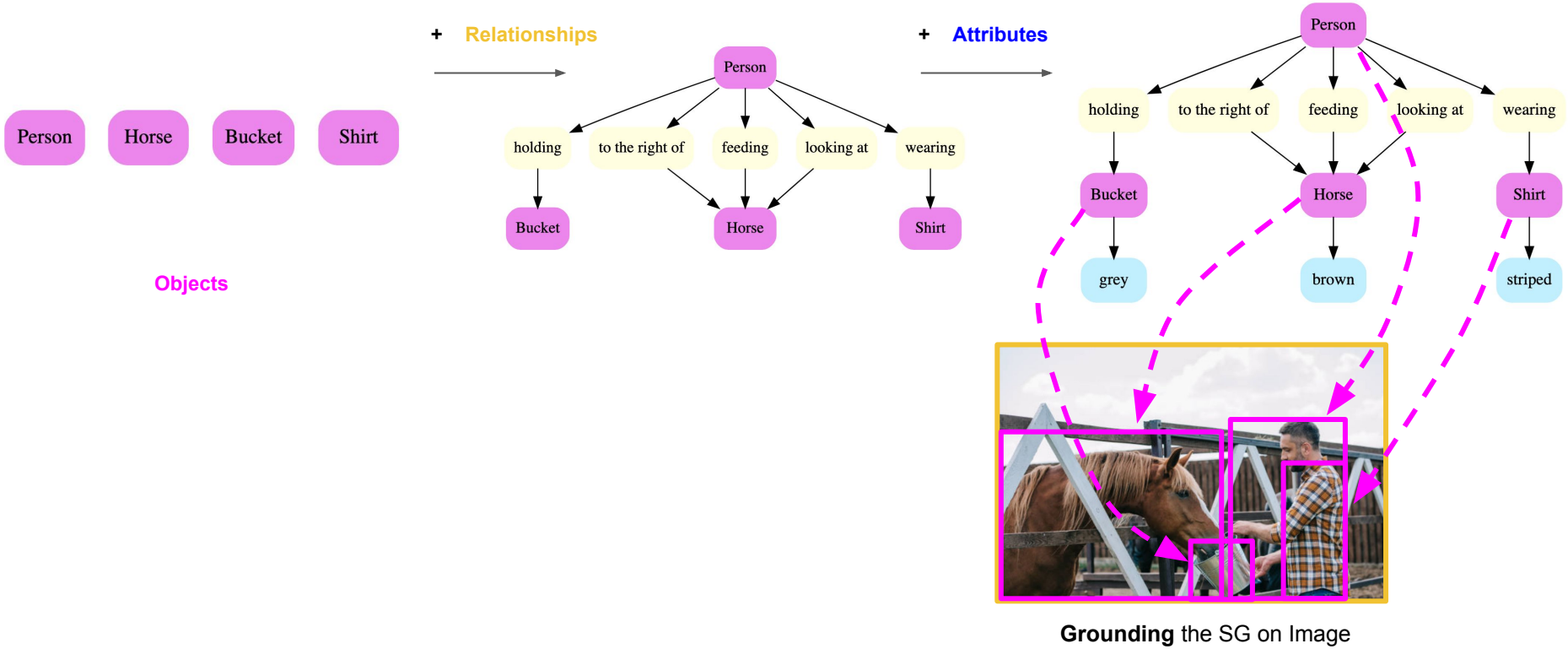
Relationships

Attributes

What is a scene graph **representation** ?

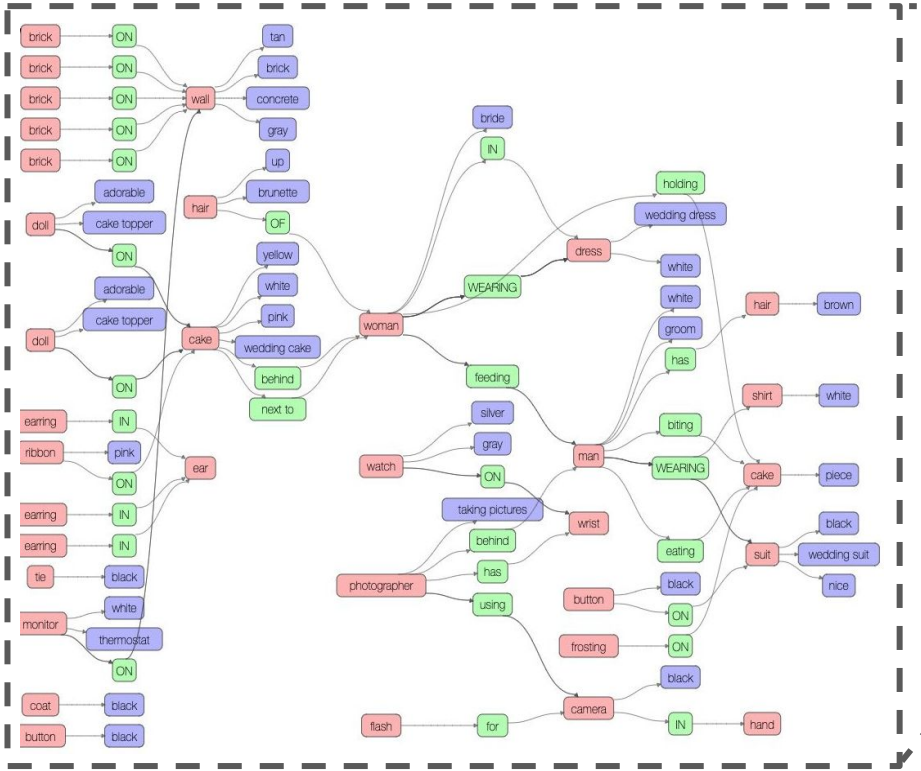


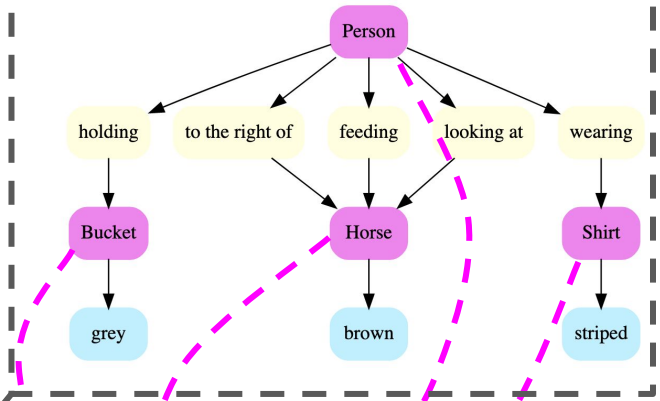
What is a scene graph **representation** ?



Grounding the SG on Image

Knowledge Graph vs Scene Graph

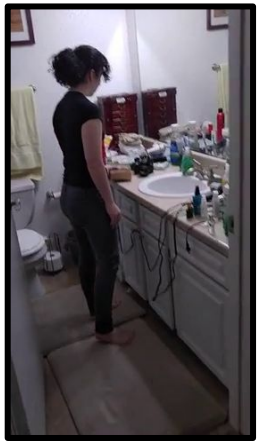




Grounding the SG on Image

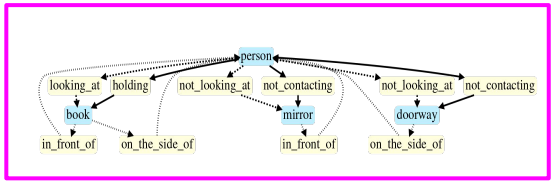
[1] Krishna, Ranjay, et al. "Visual genome: Connecting language and vision using crowdsourced dense image annotations." *International journal of computer vision* 123.1 (2017): 32-73.

Broader Context

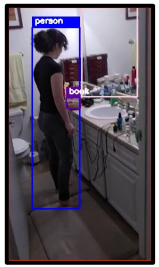


Low Semantic Value

High Semantic Value



Corners/Keypoints | **Texture/Color** | **Objects** | **Scene Graphs** | **Captions**



A person is looking at the book

Low Ground Truth Uncertainity

High Ground Truth Uncertainity

Why is SG representation **important** ?

Objects

Relationships

Attributes

Visual Question Answering

what is the man riding?

how talented is the person?

is the man going to the right of the girl?

how tall is the person?

what color is the plate?

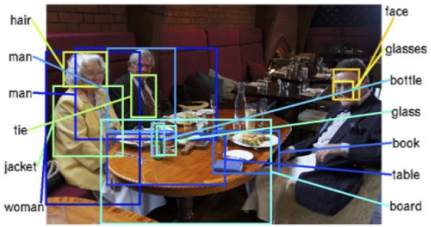
what time of day is this?

what color is the mat?

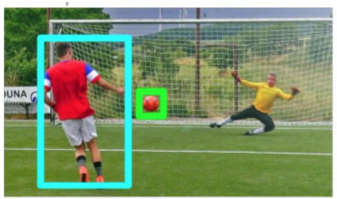
Is it day or night?

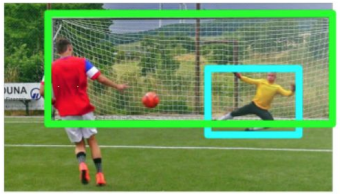
Captioning

A old sitting woman with gray hair is wearing glasses on her face and wearing a open yellow jacket. The glasses are below the thick, gray hair. She is sitting next to and is behind a wooden, brown table that has a green, glass bottle and a blue book on it. The woman is sitting next to a white, caucasian, smiling, old, sitting man wearing a striped tie.



Relationship Grounding

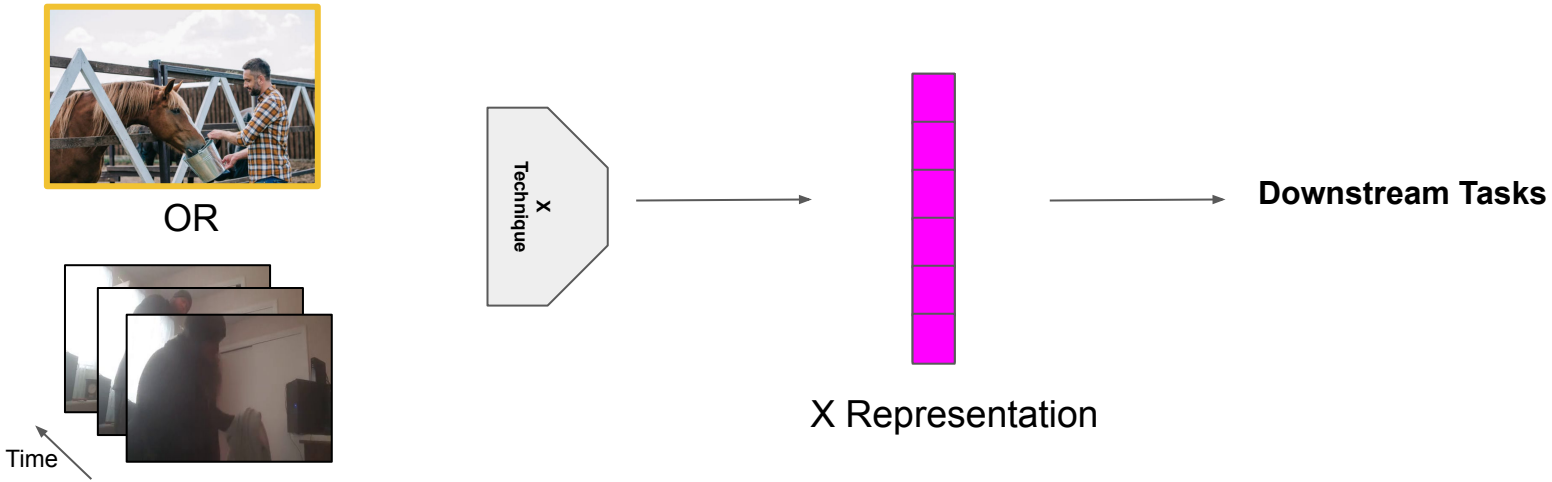




[1] Krishna, Ranjay, Michael Bernstein, and Li Fei-Fei. "Information maximizing visual question generation." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2019.
[2] Krishna, Ranjay, et al. "Referring relationships." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018
[3] Johnson, Justin, et al. "Image retrieval using scene graphs." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015

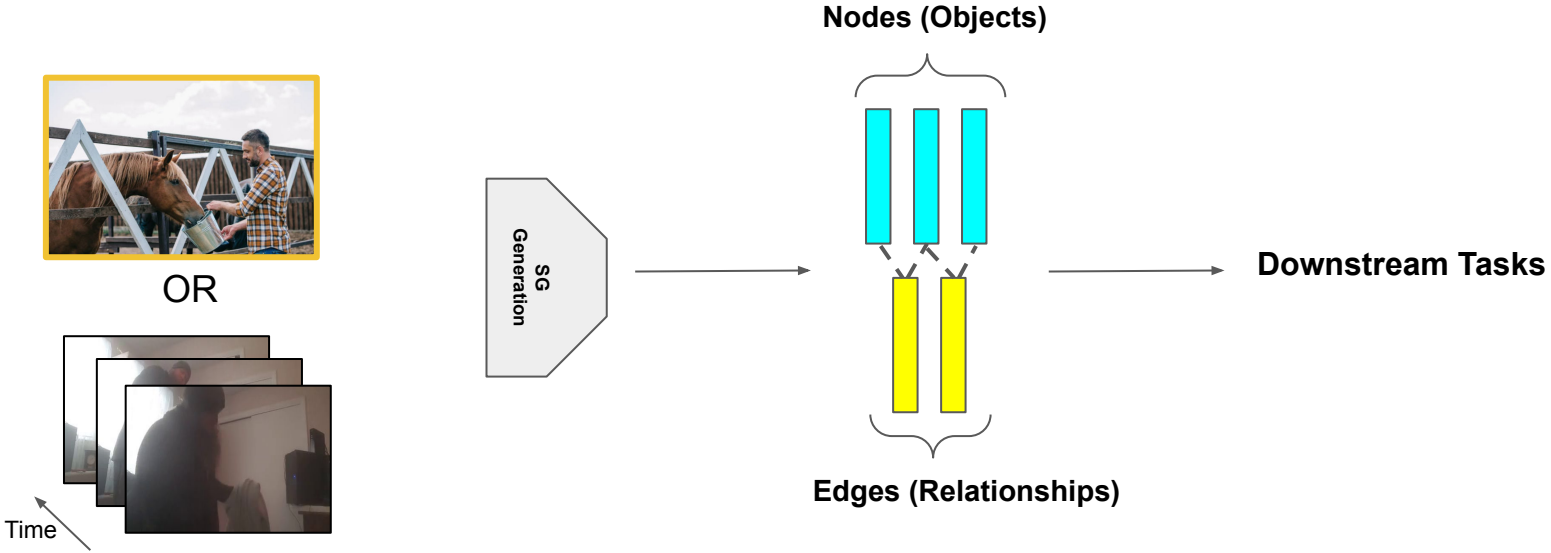
Why is SG representation **important** ?

- We want to get a **good** representation
  - Helps in downstream tasks
  - Is interpretable to humans



Why is SG representation **important** ?

- We want to get a **good** representation
  - Helps in **downstream tasks**
  - Is **interpretable** to humans



[1] Wang et al. The vqa-machine: Learning how to use existing vision algorithms to answer new questions CVPR 2017
[2] Yao et al. Exploring Visual Relationship for Image Captioning, ECCV 2018

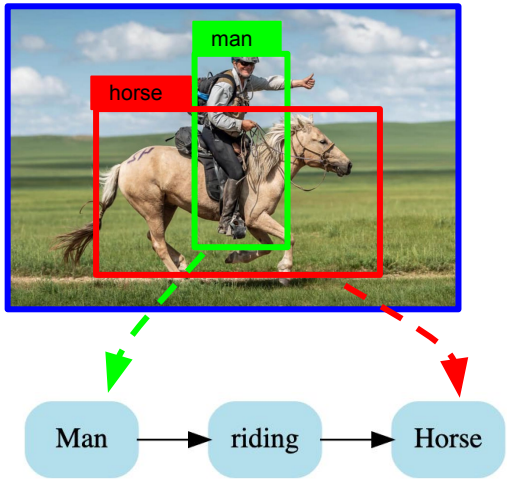
How to generate SG ?

SG Generation

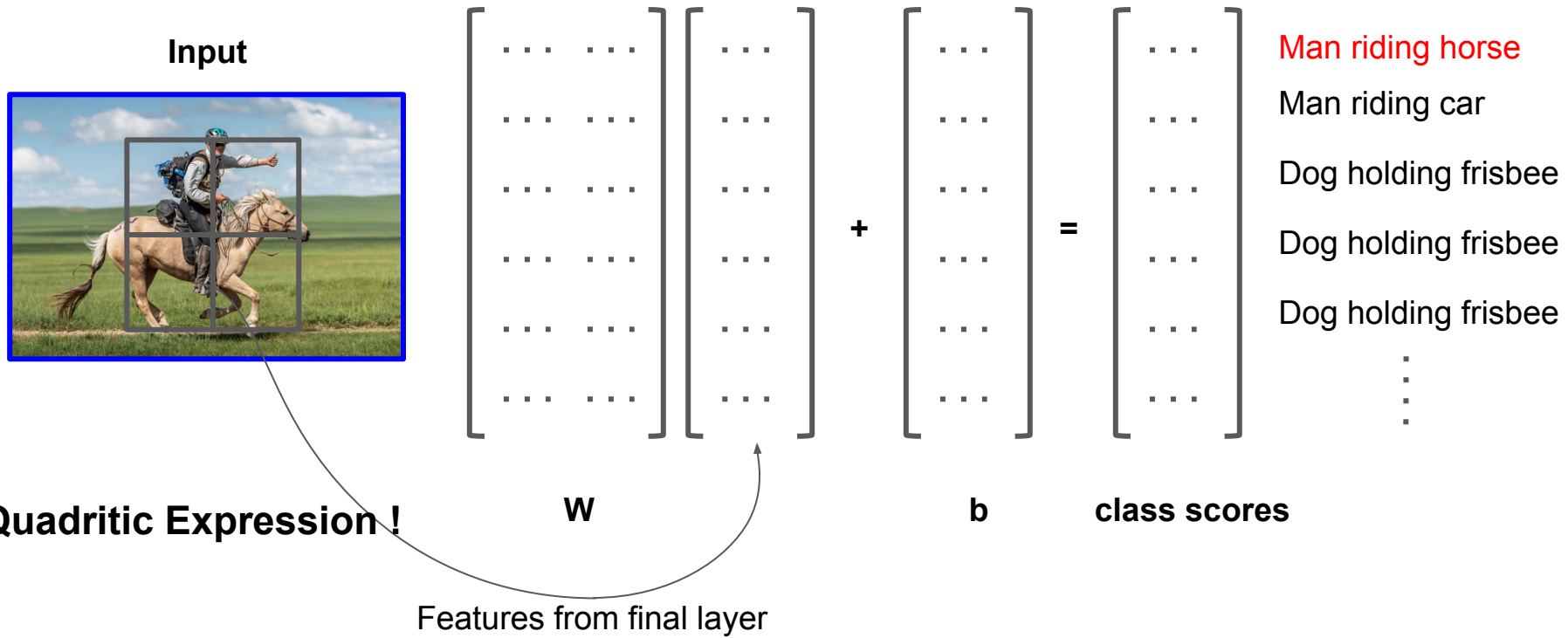
Input

SG Generation

Output



SG Generation



SG Generation

- N objects
- K relationships
- N^2K detectors
  - Not feasible

- For Visual Genome
  - N ~ 42000
  - K ~ 32000

Quadritic Expression !

SG Generation

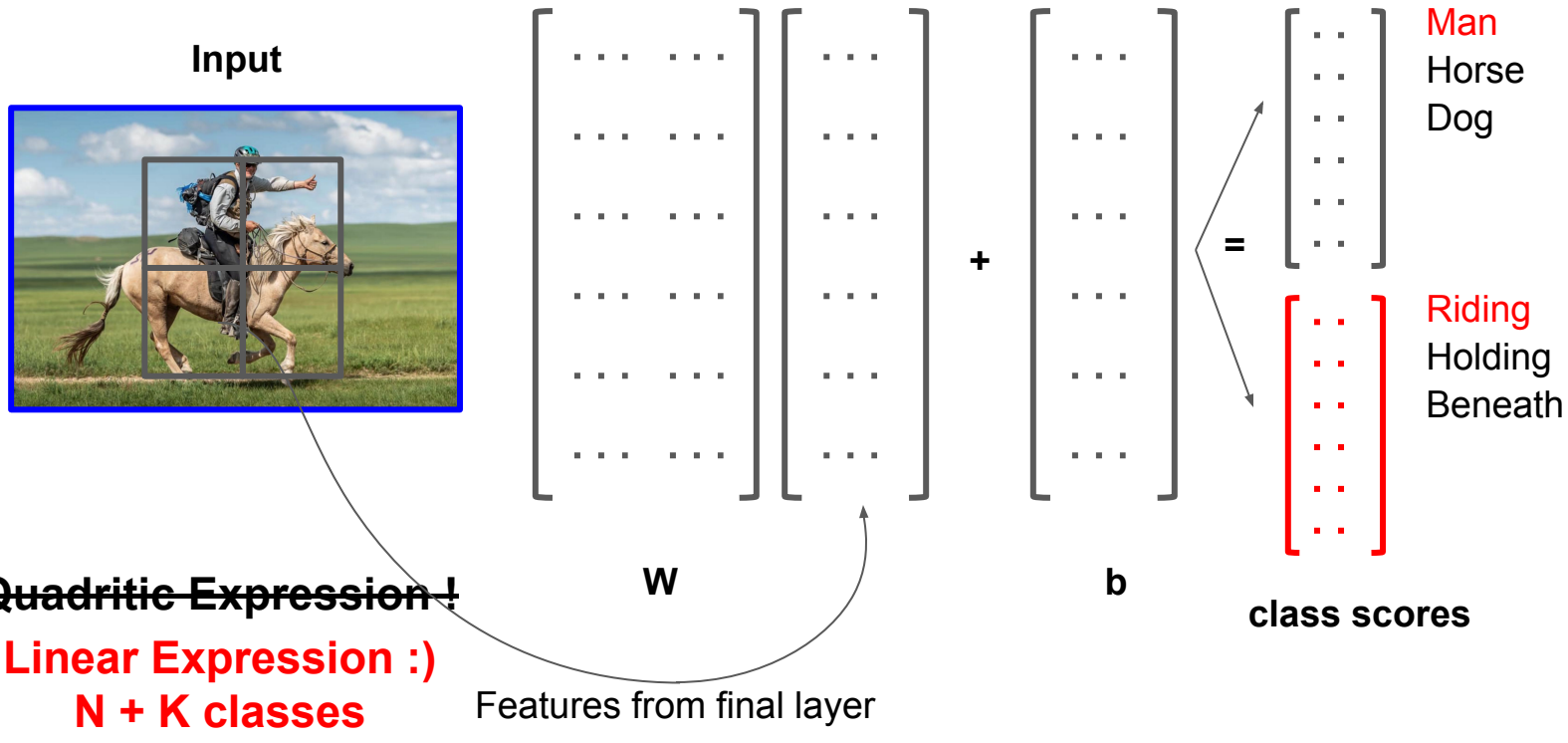
- N objects
- K relationships
- N^2K detectors
  - Not feasible

- For Visual Genome
  - N ~ 42000
  - K ~ 32000

Long Tailed Dist !!

Long Tailed Distribution

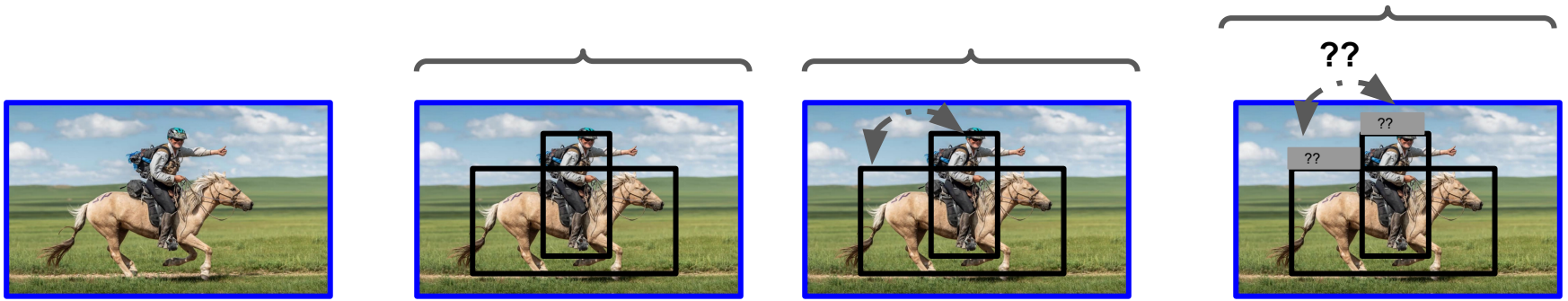
SG Generation



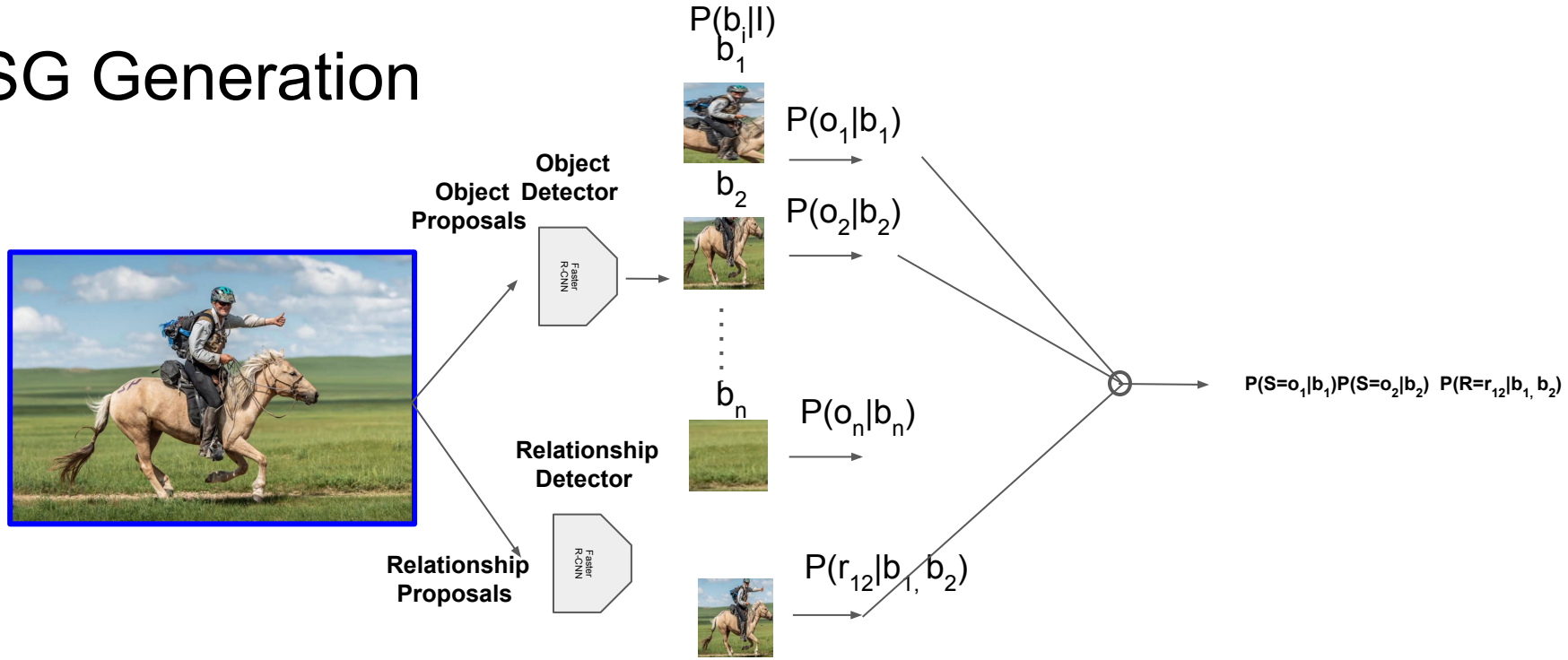
Scene Graphs

- Determine the objects and relationships between them

$$P(V, E, R, O|I) = P(V|I)P(E|V, I)P(R, O|E, V, I)$$

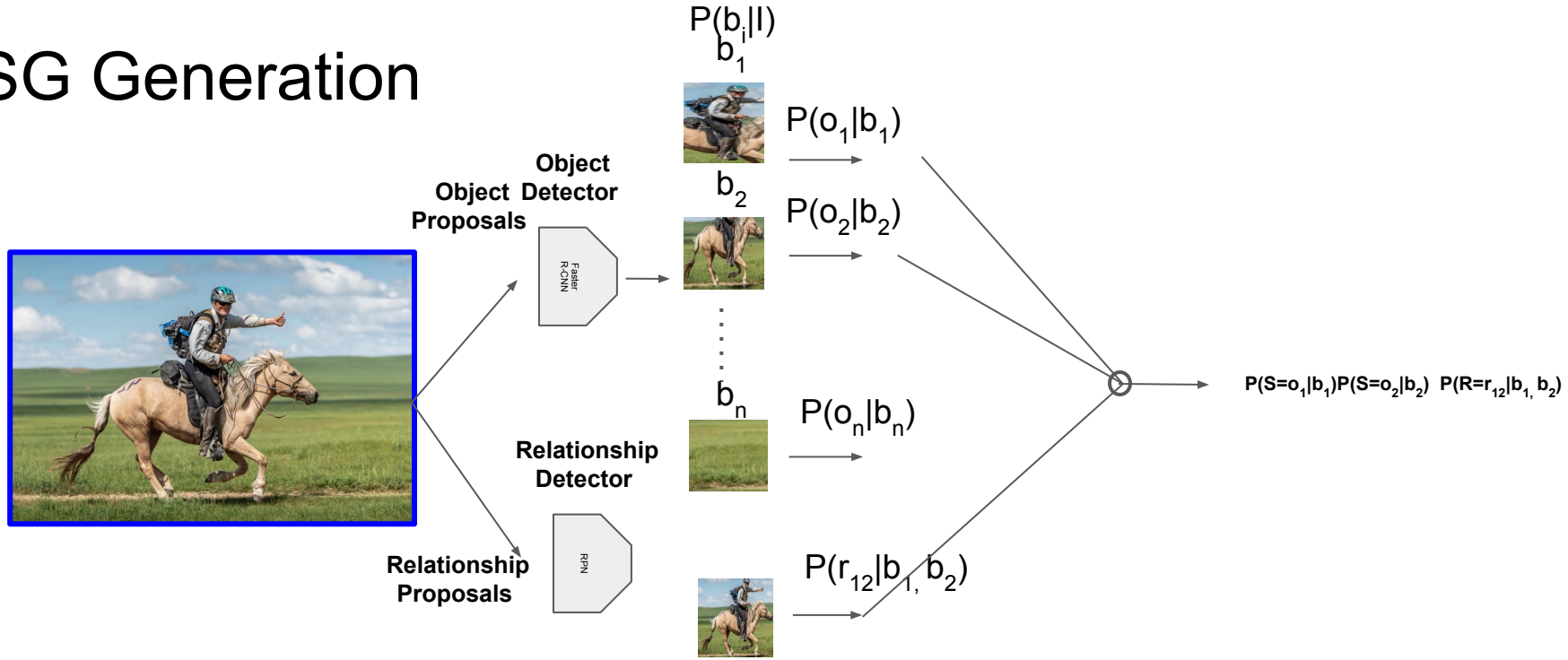


SG Generation



$$\mathbf{P(S=o_1,O=o_2,R=r_{12}|I)} = P(B=b_1|I)P(B=b_2|I)\ \ P(S=o_1|b_1)P(S=o_2|b_2)\ \ P(R=r_{12}|b_1, b_2)$$

SG Generation

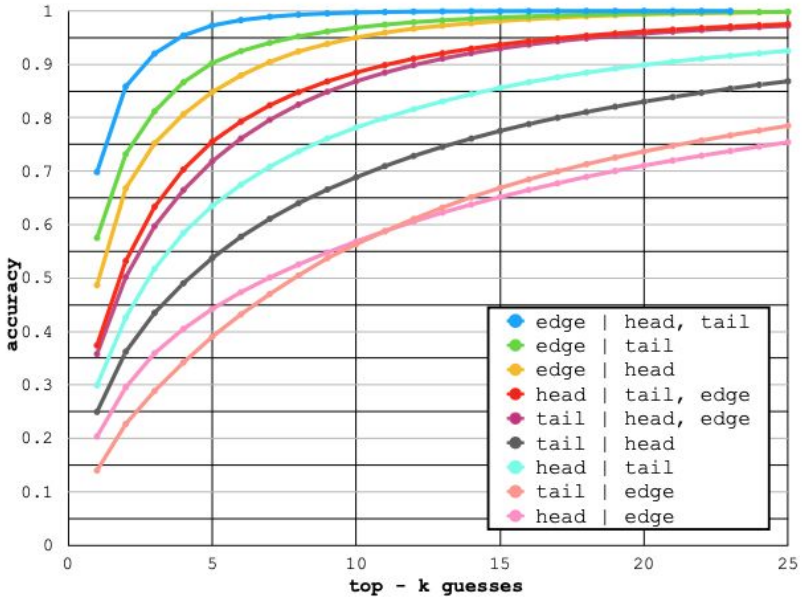


$$\mathbf{P(S=o_1,O=o_2,R=r_{12}|I)} = P(B=b_1|I)P(B=b_2|I)\ \ P(S=o_1|b_1)P(S=o_2|b_2)\ \ P(R=r_{12}|b_1, b_2)$$

$$\mathbf{argmax\ P(S=o_1,O=o_2,R=r_{12}|I)}$$

Frequency Based

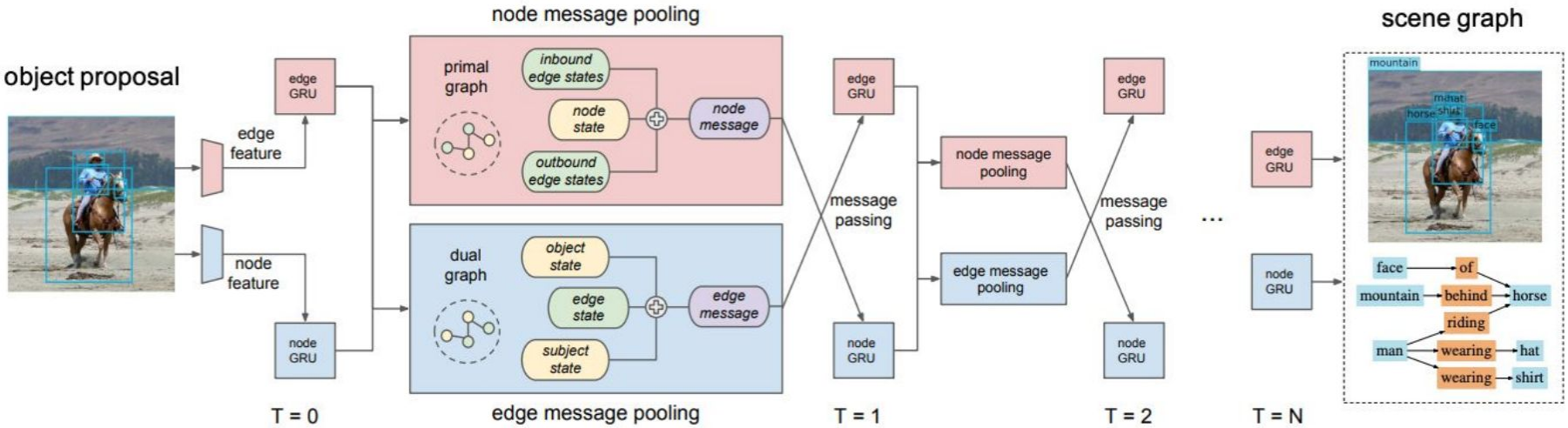
- Visual Genome



[1] Zellers, Rowan, et al. "Neural motifs: Scene graph parsing with global context." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018.

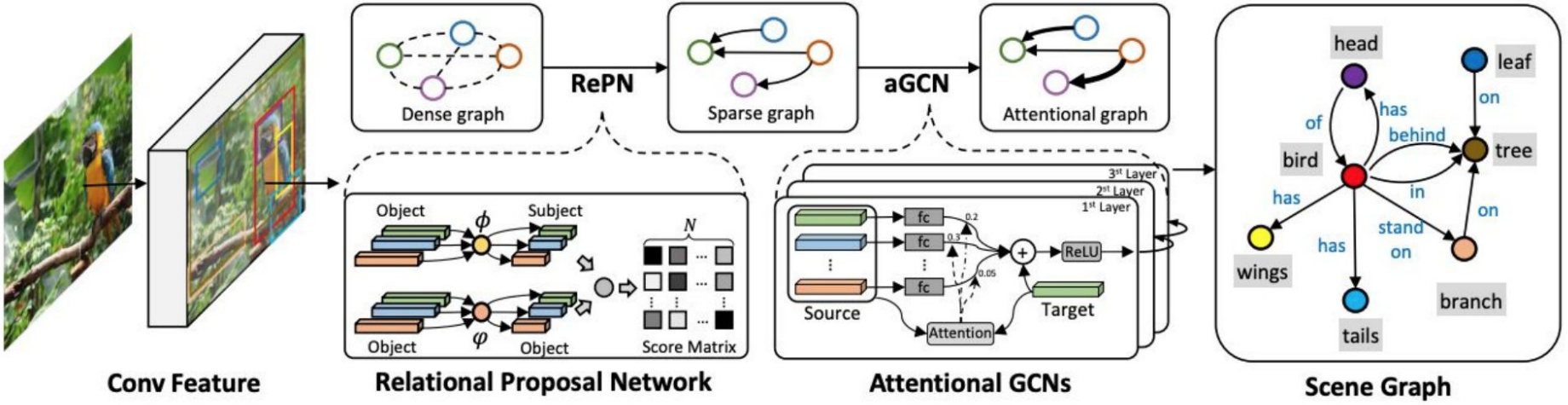
Iterative Message Passing [IMP]

- Use the bipartiteness of SG



Graph R-CNN

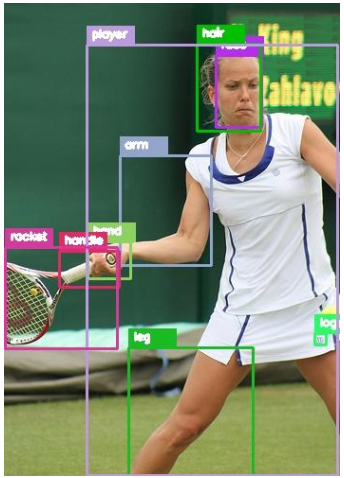
- Extract a sparse graph
- Use attentional GCN

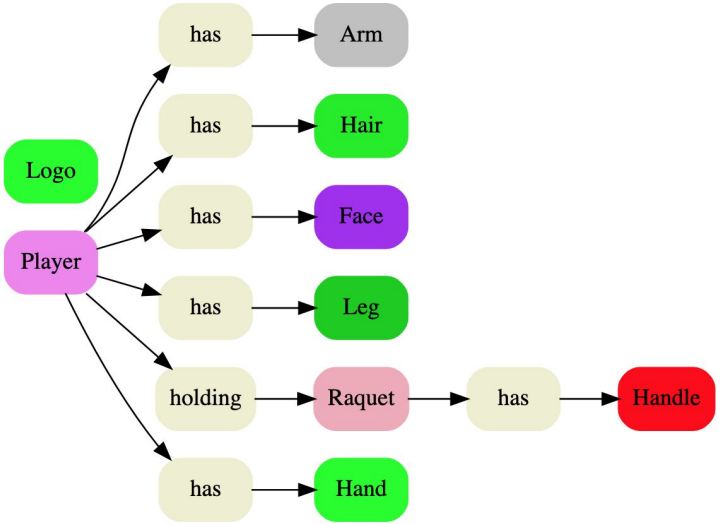


Graph R-CNN [G-RCNN]

Evaluation

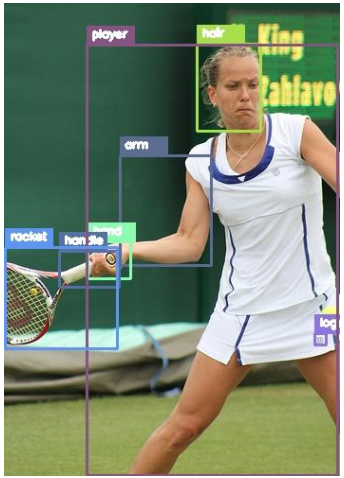
- Minimum Graph Edit Distance [GED]
  - Insert/Delete a vertex
  - Insert/Delete an edge
  - Change the label of a vertex

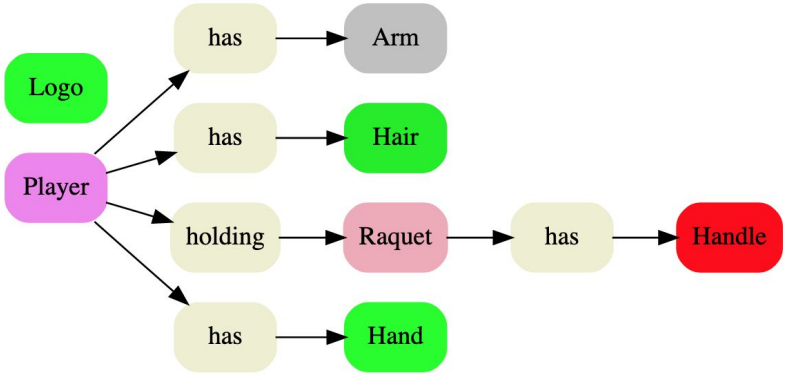




Evaluation

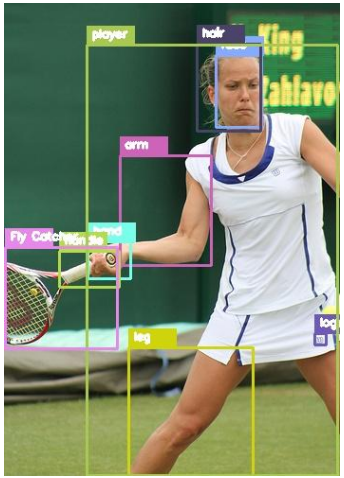
- Minimum Graph Edit Distance [GED]
  - **Insert/Delete a vertex**
  - **Insert/Delete an edge**
  - Change the label of a vertex

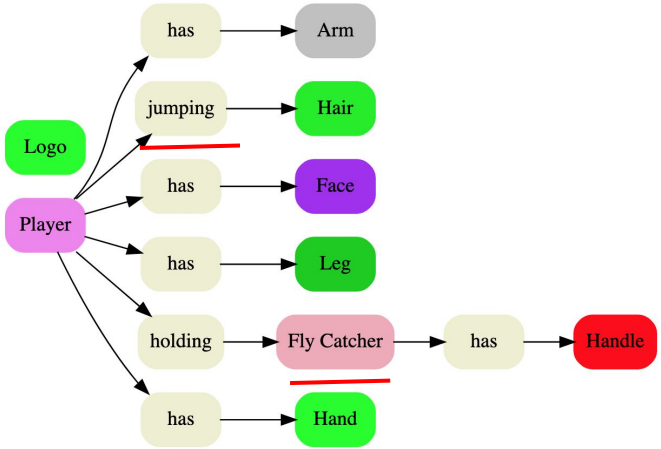




Evaluation

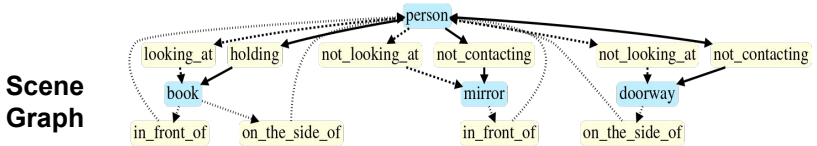
- Minimum Graph Edit Distance [GED]
  - Insert/Delete a vertex
  - Insert/Delete an edge
  - **Change the label of a vertex**





Evaluation

- NP-Hard or APX-hard
  - Use a triplet-recall based method



$[(x_{p_1}, y_{p_1}), (x_{p_2}, y_{p_2})]_{GT}$

GT Triplets

person looking at book
person holding book
person not looking at mirror
person not contacting mirror
person not looking at doorway
person not contacting doorway
doorway on the side of person
mirror in front of person
book in front of person
book on the side of person

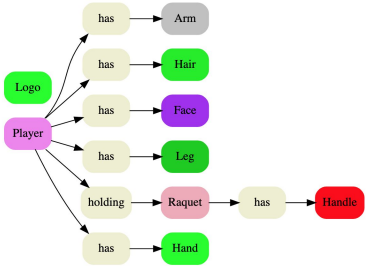
$[(x_{p_1}, y_{p_1}), (x_{p_2}, y_{p_2})]_{Pred}$

Pred. Triplets (256)

person beside mirror
person leaning on mirror
person not looking at mirror
person not contacting mirror
……
……..
……
……..
……
……..
……
……

Evaluation

GT



Player has arm
Player has hair
Player has face
Player has leg
Player holding racquet
Racquet has handle
Player has hand

Player has arm	0.98
Dog has hair	0.88
Player has face	0.76
Player wearing shirt	0.74
Women holding racquet	0.73
Racquet has handle	…
Player has hand	…
….	…
….	…
….	…

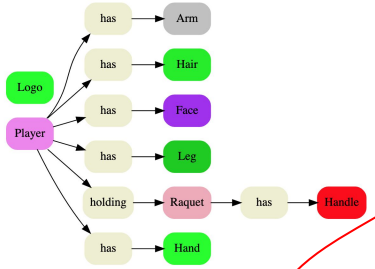
Predicted

256 Triplets

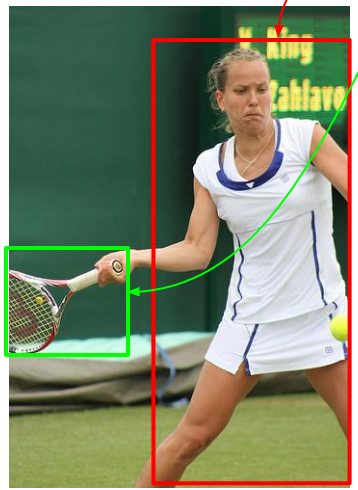
SG Detection

Evaluation

GT



Player has arm
Player has hair
Player has face
Player has leg
Player holding racquet
Racquet has handle
Player has hand



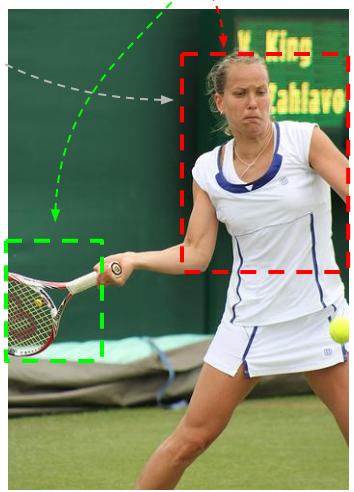
IoU > 0.5

SG Detection

Player has arm	0.98
Dog has hair	0.88
Player has face	0.76
Player wearing shirt	0.74
Player holding racquet	0.73
Racquet has handle	…
Player has hand	…
….	…
….	…
….	…

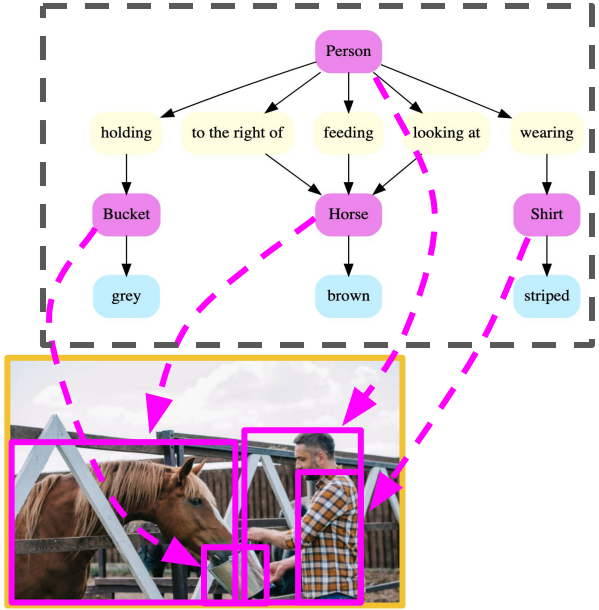
Predicted

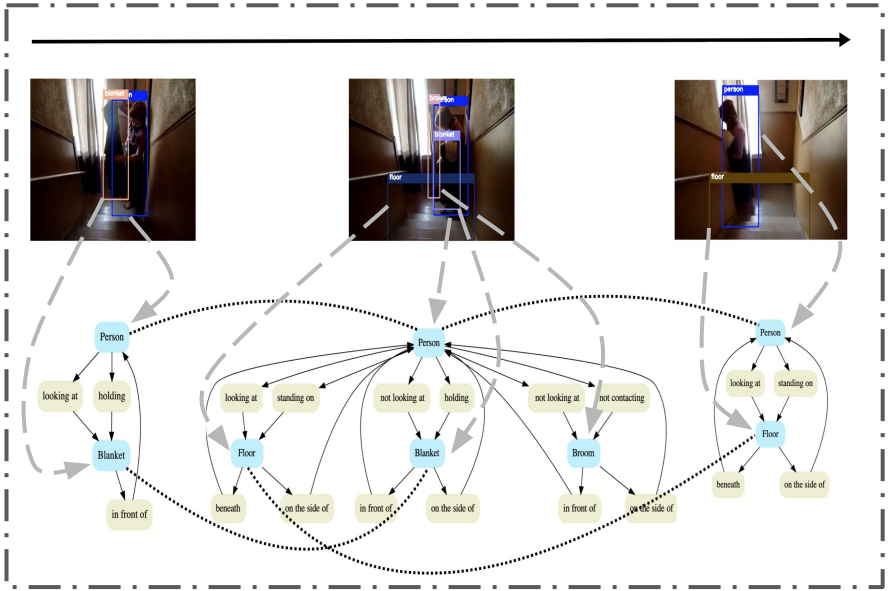
256 Triplets



From Images to Videos

From Images to Videos



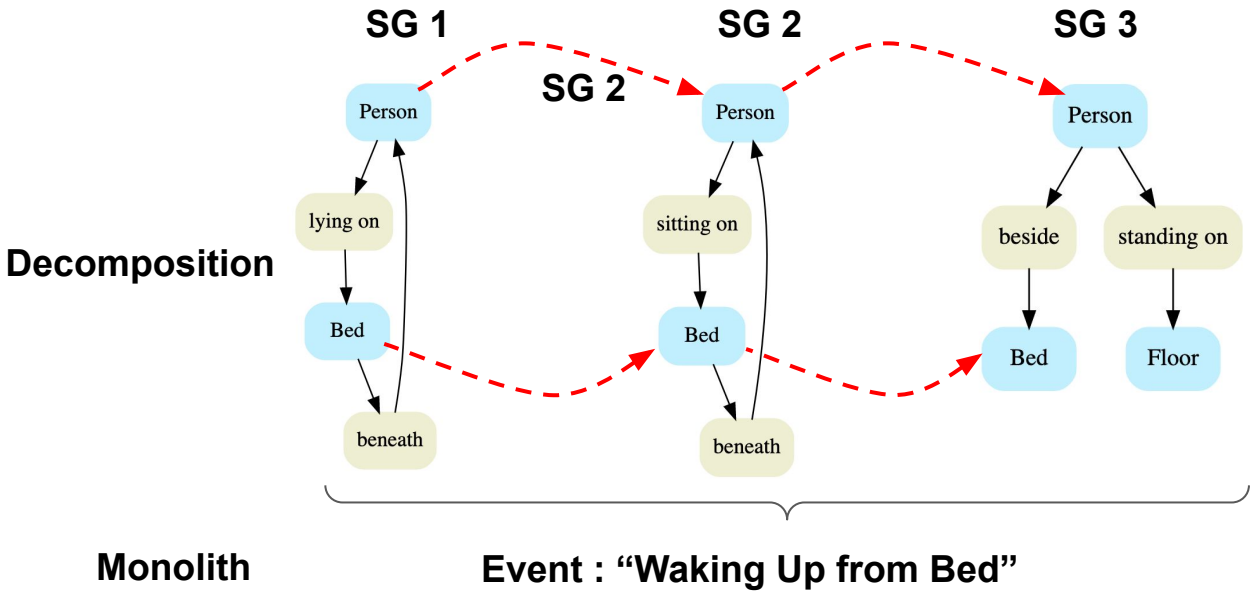


Scene Graphs

Spatio-temporal Scene Graph

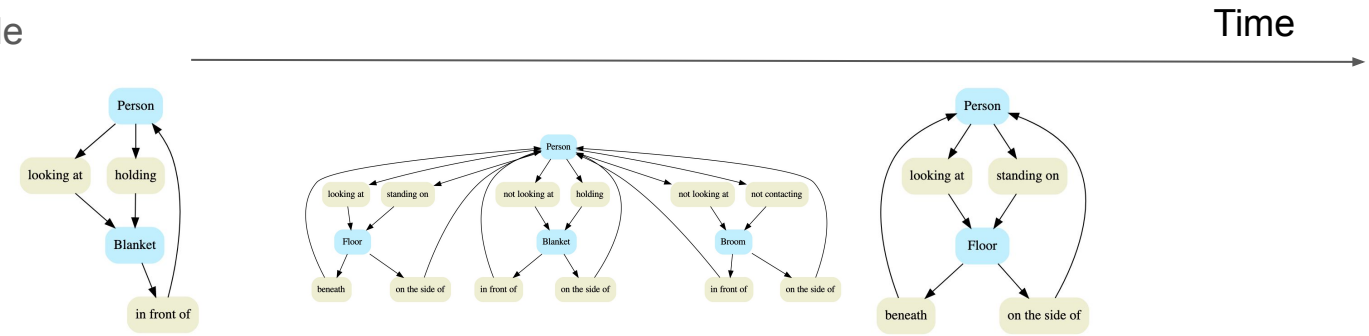
From Images to Videos

- Activities as partonomies
  - Like objects, activities can be divided into parts
  - Events are made of action-object interaction



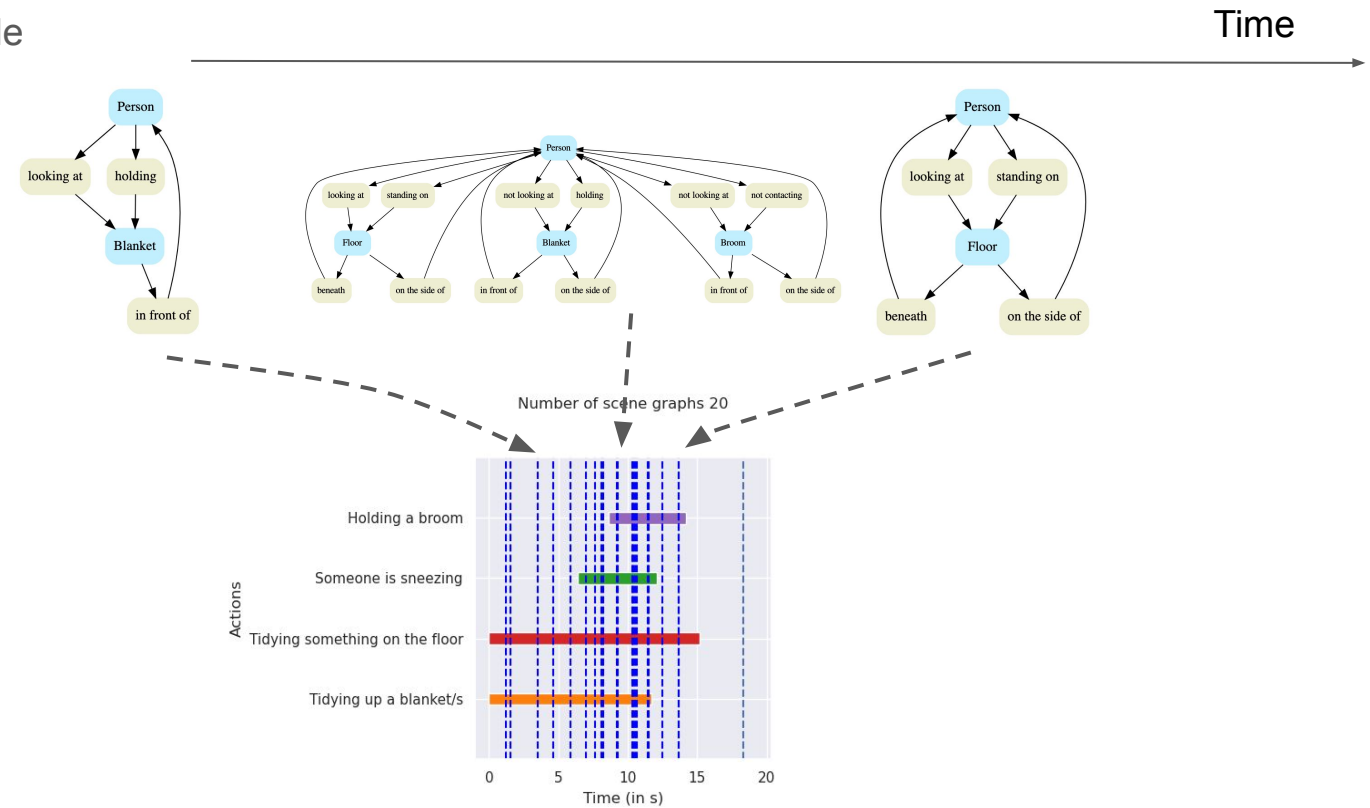
Why are Spatio-temporal SG important ?

- Puzzle

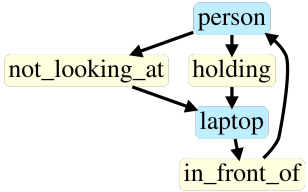


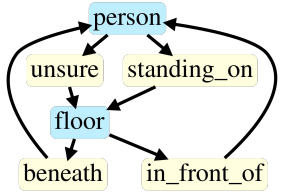
Why are Spatio-temporal SG important ?

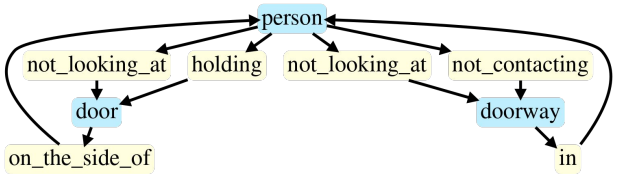
- Puzzle



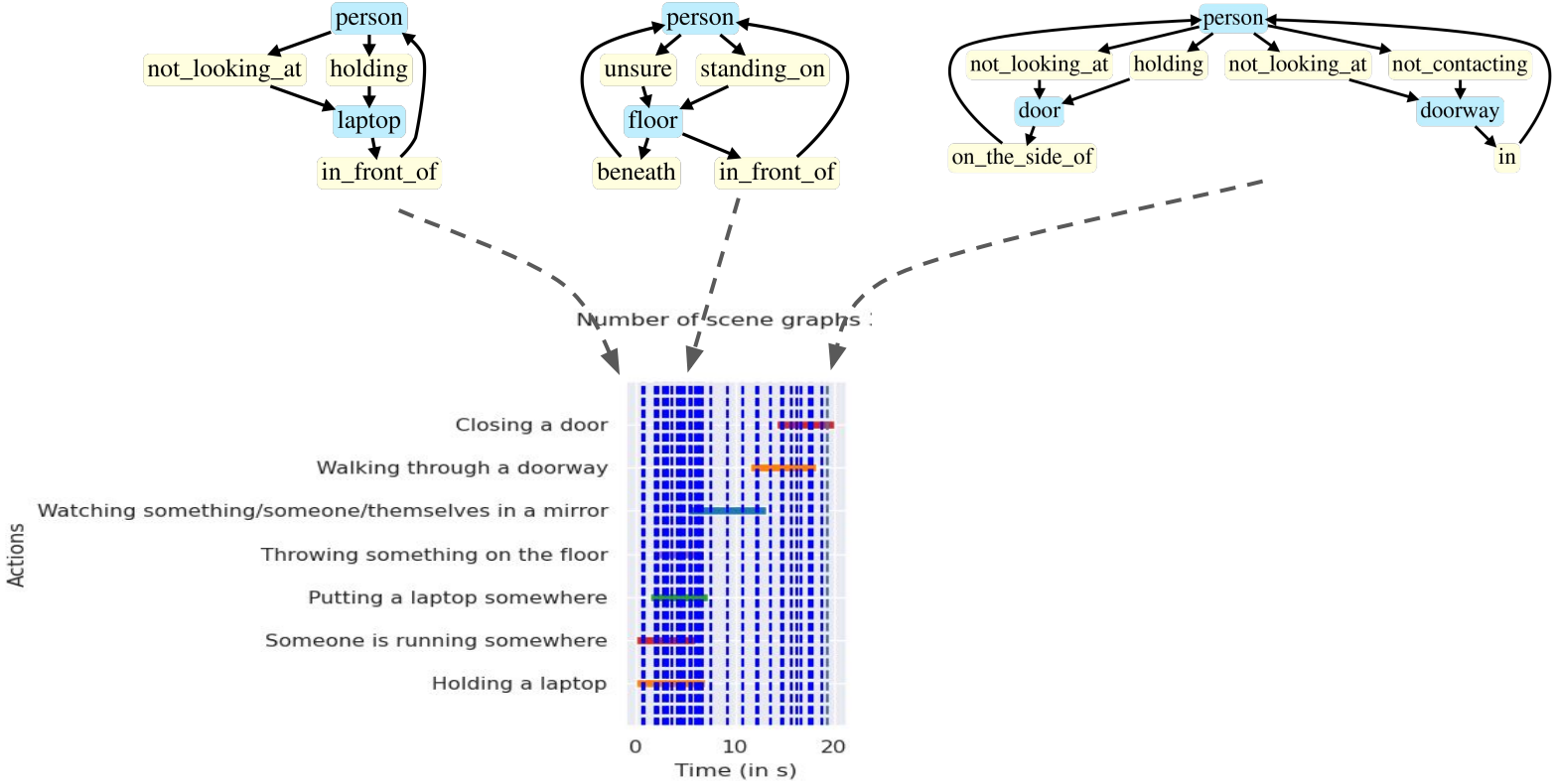
Why are Spatio-temporal SG important ?





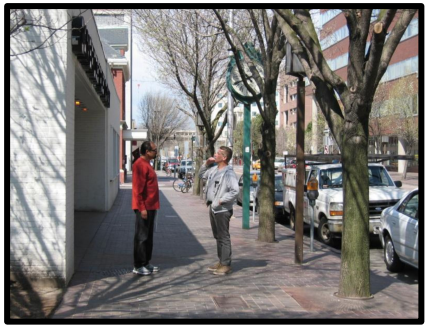


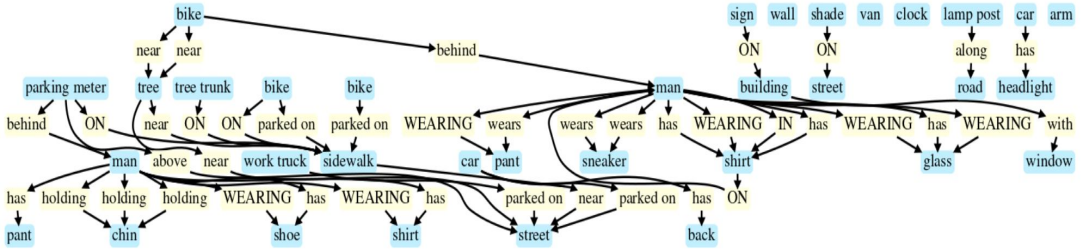
Why are Spatio-temporal SG important ?



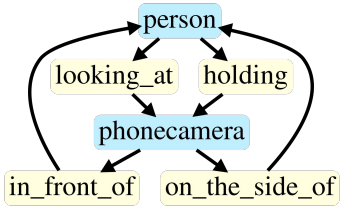
VG [Images] to AG[Videos]

- Visual Genome



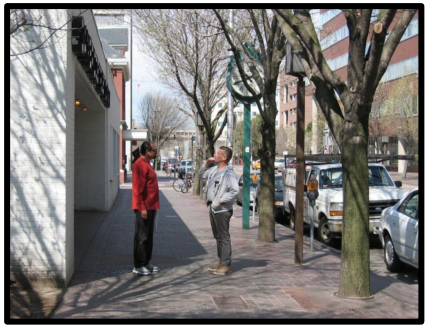


- Action Genome



VG [Images] to AG[Videos]

- Visual Genome

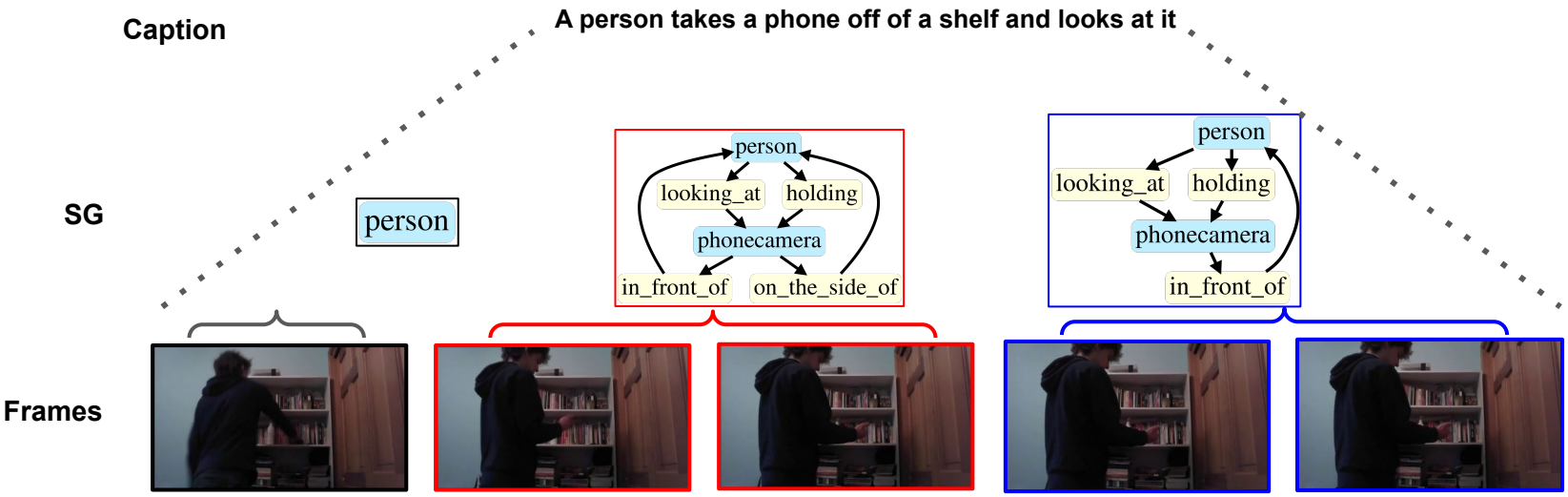


- Dense annotation

- Action Genome

- Only objects involved in an event are annotated
- Only one person is annotated
  - In case multiple people exist in the video

Another peculiarity



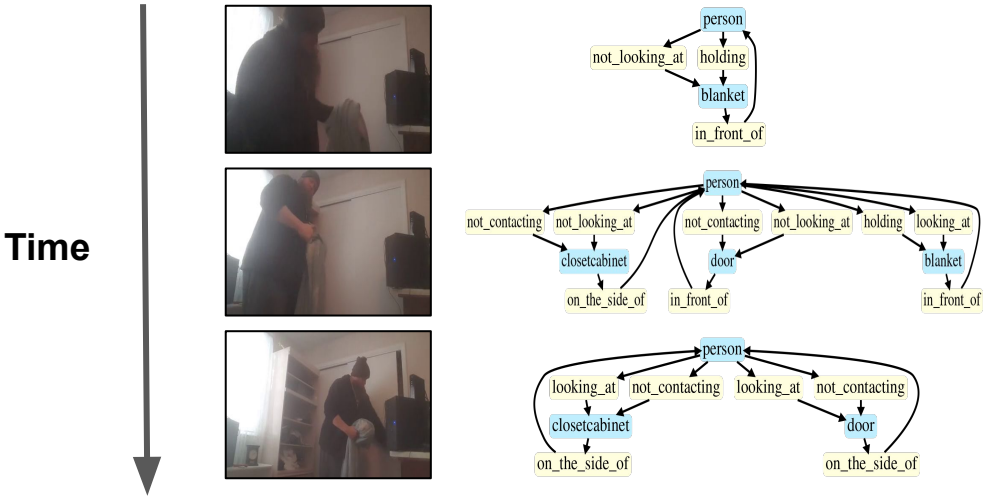
Changes **more** frequently

Changes **less** frequently

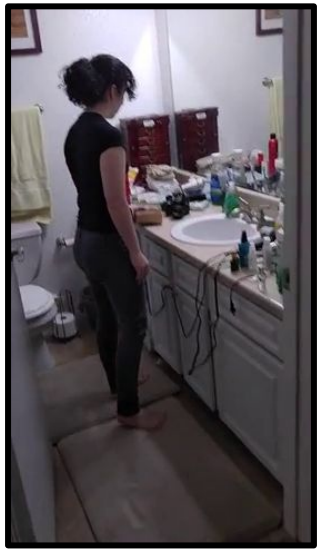
Corners/Keypoints | Texture/Color | Objects | Scene Graphs | Captions

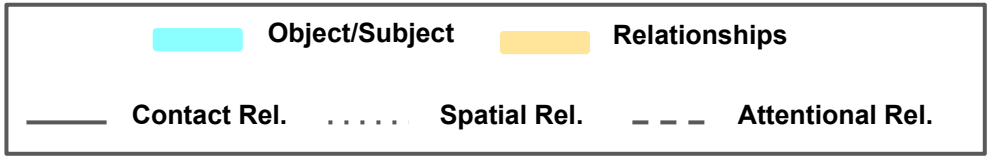
Video SG Generation

- **Hypothesis** - Predicting SG into the future will help understand the actions performed in the videos
  - The *relationships* in a SG correspond to the actions themselves
- **Evolution** of a SG hints at the action performed within a video

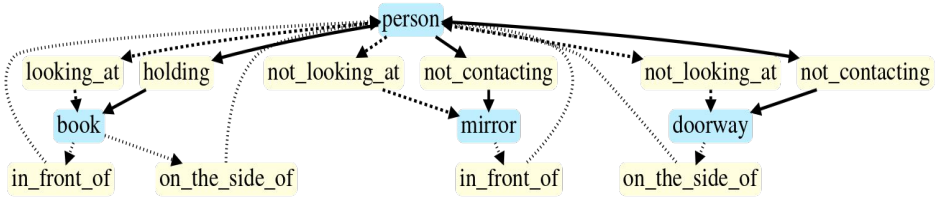


Action Genome - Analysis





Scene Graph

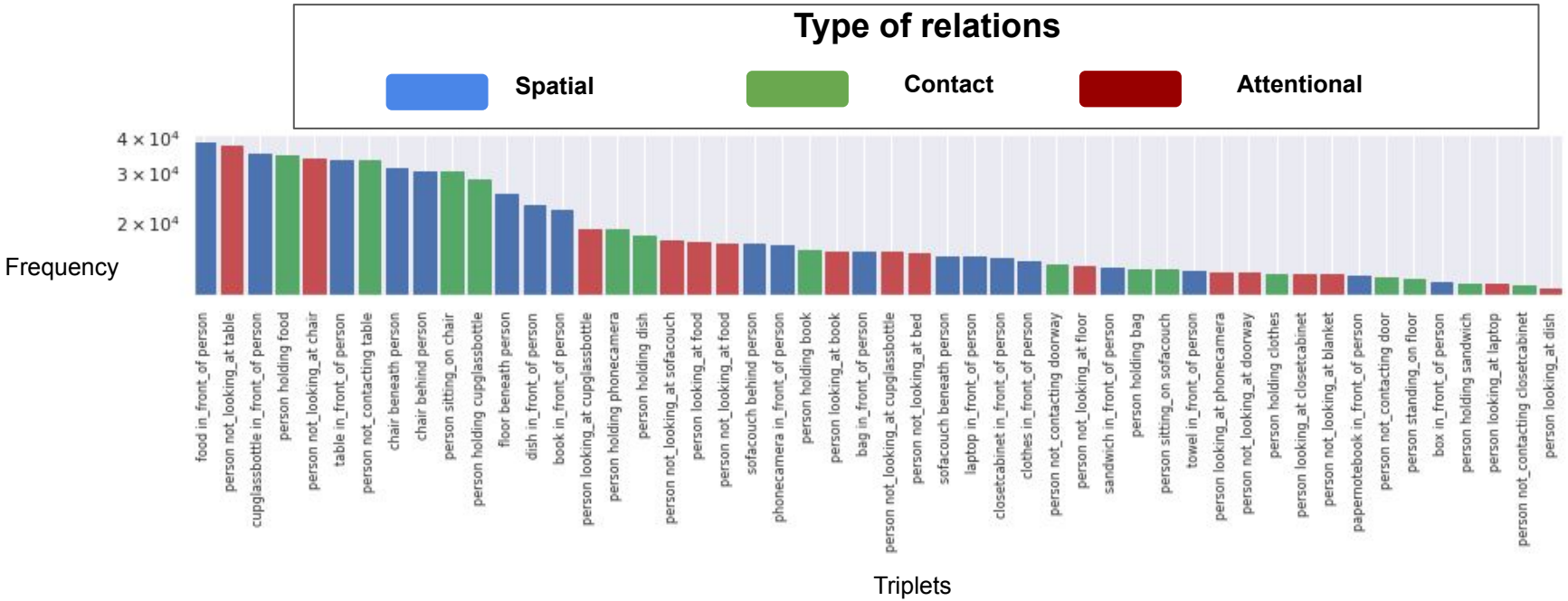


Triplets

person looking at book
person holding book
person not looking at mirror
person not contacting mirror
person not looking at doorway
person not contacting doorway
doorway on the side of person
mirror in front of person
book in front of person
book on the side of person

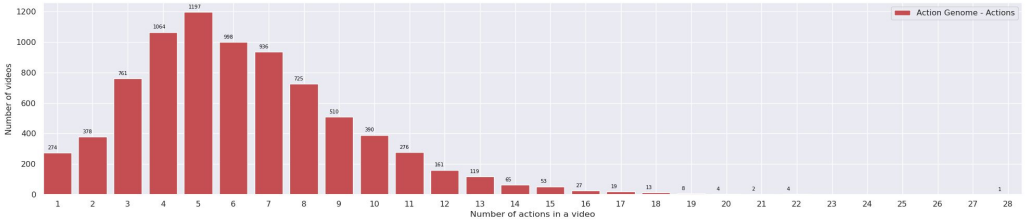
Action Genome - Triplets

- <Subject., Relationship., Object.> Triplets

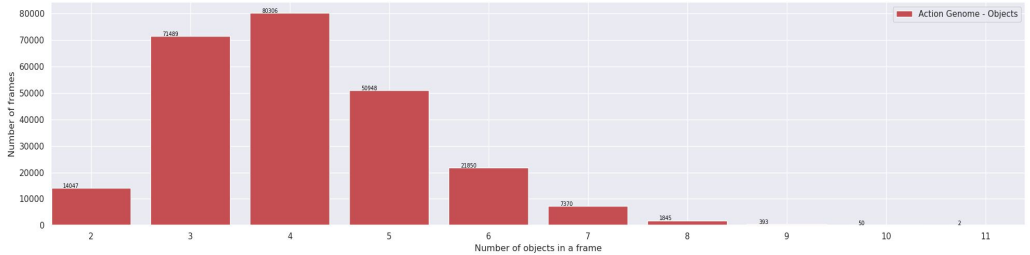


Action Genome

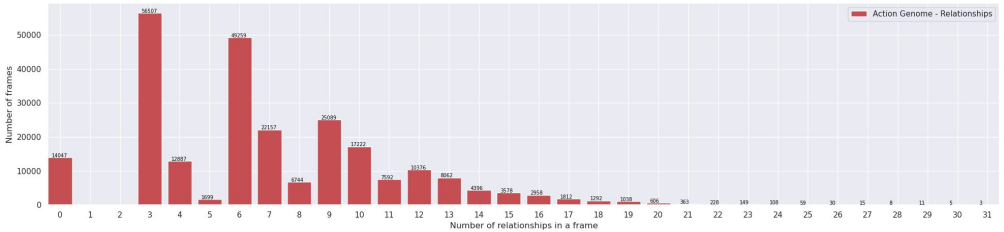
- Number of actions in a video



- Number of objects in a frame

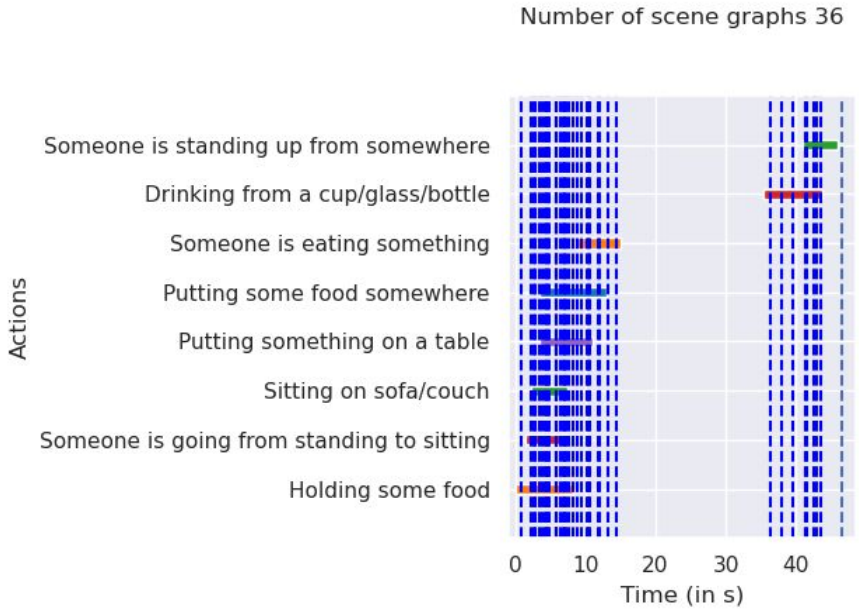


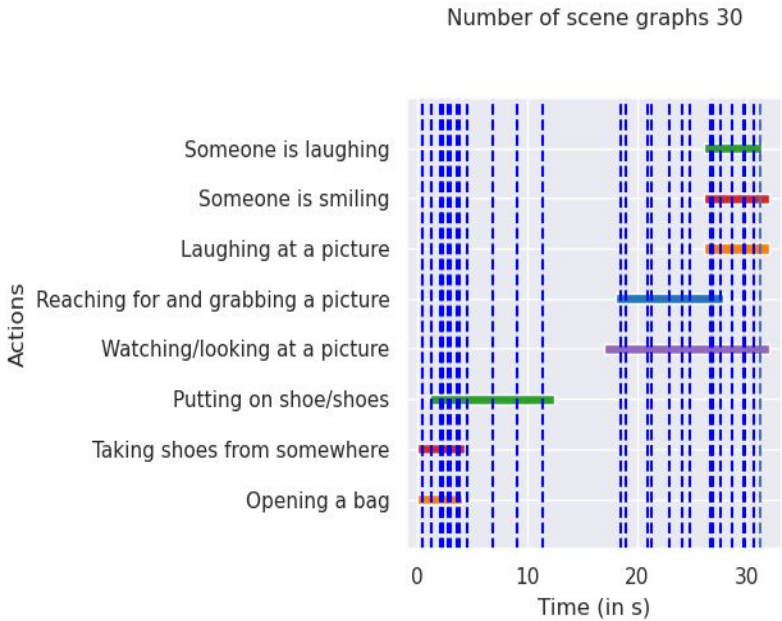
- Number of rel in a SG



Action Genome - Gantt Charts

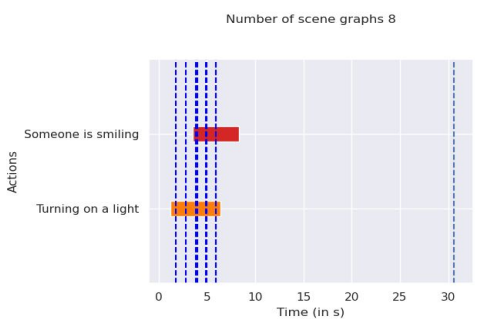
- Frames annotated with SG
  - **5 per action - sampled uniformly in action interval**

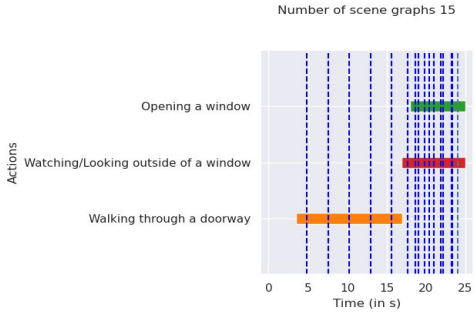


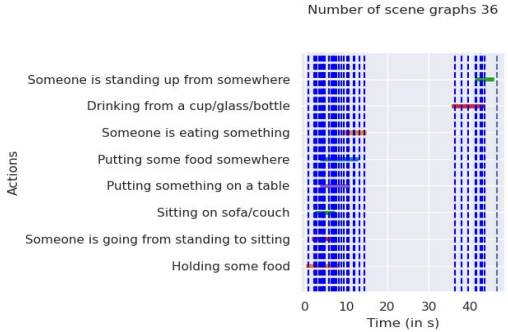


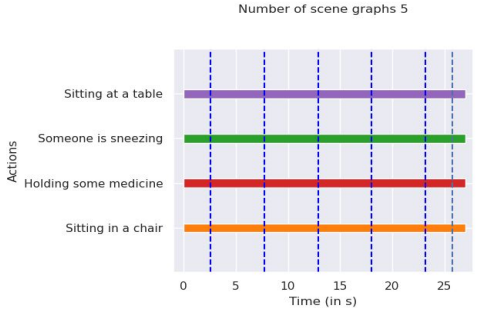
Action Genome - Gantt Charts

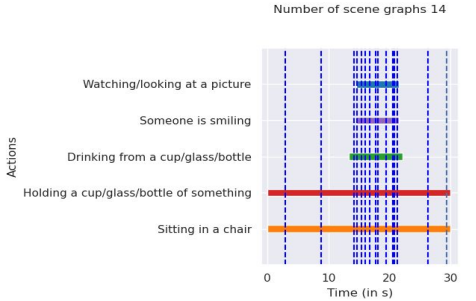
- Sample density proportional to actions

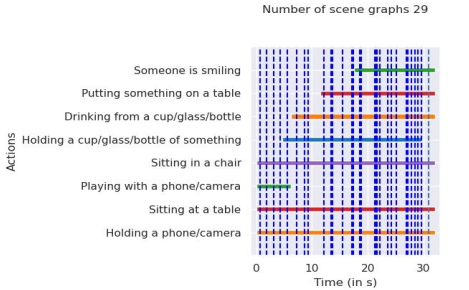




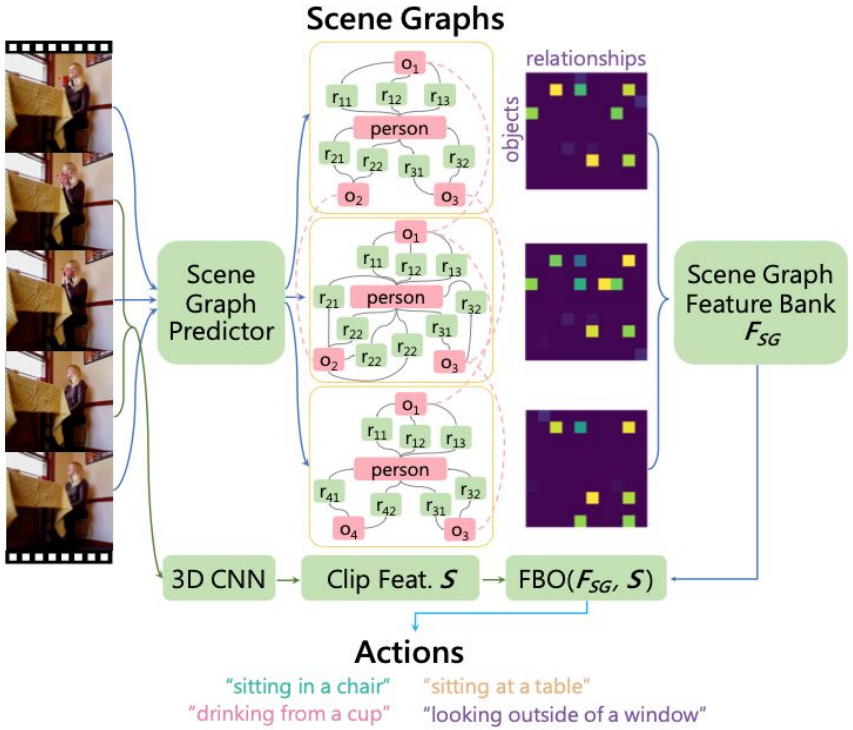








Action Genome - Baseline Exp.



Experiments

- Comparison with AG Paper [1]
- Comparison with papers reporting results on VG
  - IMP [2]
  - MSDN [3]
  - G-RCNN [4]

	Papers			Reproduced		
Method	SGDet@20	SGDet@50	SGDet@100	SGDet@20	SGDet@50	SGDet@100
Baseline	12.3	15.8	17.7	-	-	-
MSDN	19.2	23.9	26.3	17.62	22.35	25.07
IMP	19.2	23.8	26.2	18.85	23.43	25.92
G-RCNN	-	-	-	17.26	22.3	25.44
RELDN	-	-	-	20.12	24.78	26.96

Results for Visual Genome Dataset

[1] Ji, Krishna et al. Action Genome: Actions as Compositions of Spatio-Temporal Scene Graphs, CVPR 2020
[2] Xu, Danfei, et al. "Scene graph generation by iterative message passing." CVPR 2017
[3]
[4] Yang, Jianwei, et al. "Graph r-cnn for scene graph generation ECCV 2018

Experiments

- Comparison with AG Paper [1]
- Comparison with papers reporting results on VG
  - IMP [2]
  - MSDN [3]
  - G-RCNN [4]

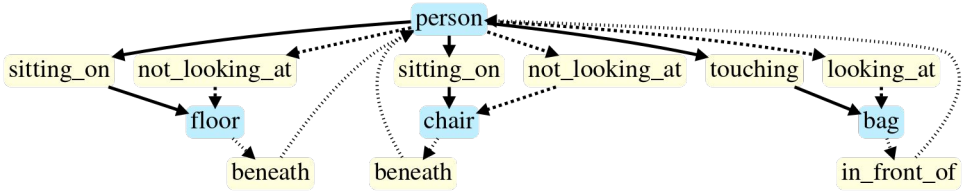
Method	Papers		Reproduced		
	SGDet@20	SGDet@50	SGDet@20	SGDet@50	SGDet@100
Freq. Based	24.03	24.87	8.87	9.51	9.6
MSDN	24.00	25.64	22.22	24.44	24.78
IMP	23.88	25.52	22.49	24.6	24.94
G-RCNN	24.12	25.77	20.07	21.94	22.25
RELDN	25.00	26.21	-	-	-

Results for Action Genome Dataset

[1] Ji, Krishna et al. Action Genome: Actions as Compositions of Spatio-Temporal Scene Graphs, CVPR 2020
[2] Xu, Danfei, et al. "Scene graph generation by iterative message passing." CVPR 2017
[3]
[4] Yang, Jianwei, et al. "Graph r-cnn for scene graph generation ECCV 2018

Challenges with AG

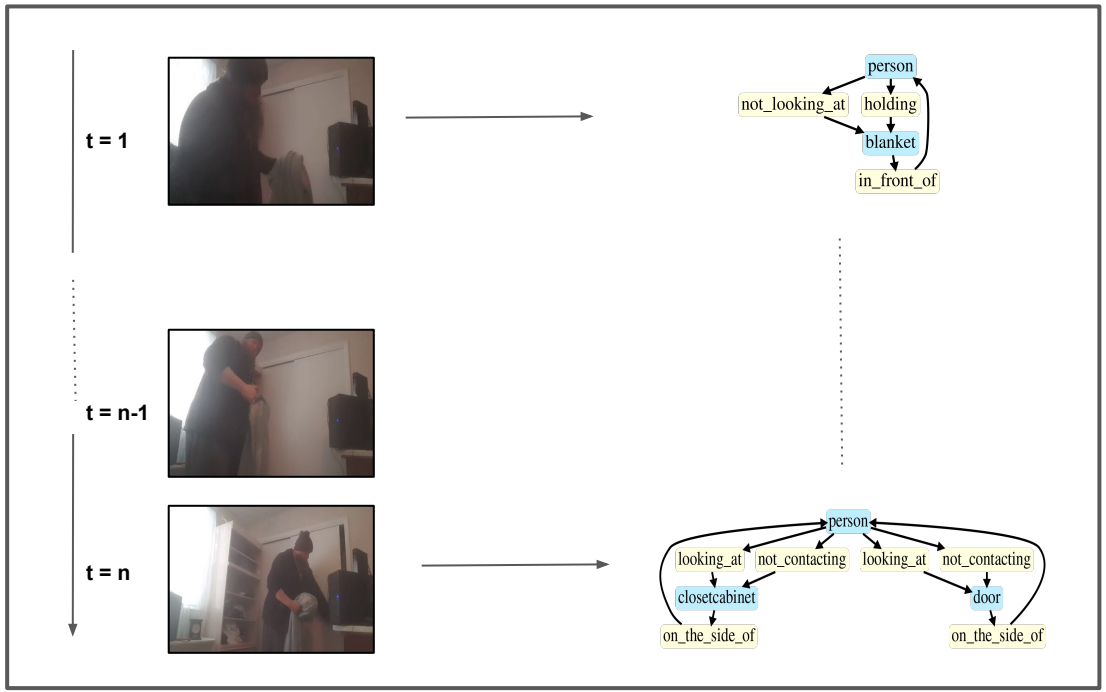
- Noisy annotations



- Directionality of relations
  - Perform experiments without spatial relations
- Annotations based on Annotations

Ideas

Formuation

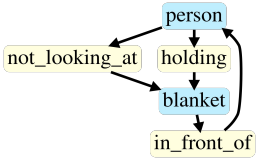

t = 1
t = n-1
t = n
person
not_looking_at
holding
blanket
in_front_of
person
looking_at
not_contacting
looking_at
not_contacting
closetcabinet
door
on_the_side_of
on_the_side_of

Formuation -- Simpler

Current Frame

Next Frame

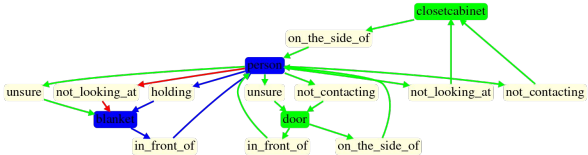
Scene Graph 1



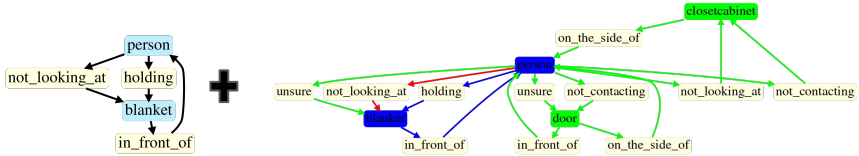
Scene Graph 2

??

Predict the diff graph



Scene Graph 2



Diff Graphs

Added relationships

Remaining relationships

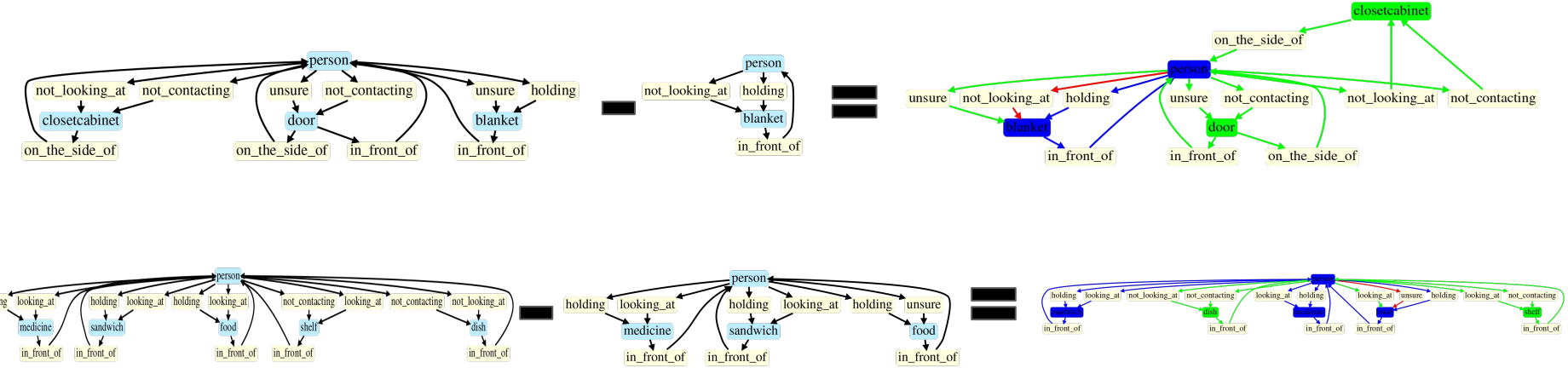
Removed relationships

Added objects

Remaining objects

Removed objects

Predicates



Diff Graphs

Added relationships

Remaining relationships

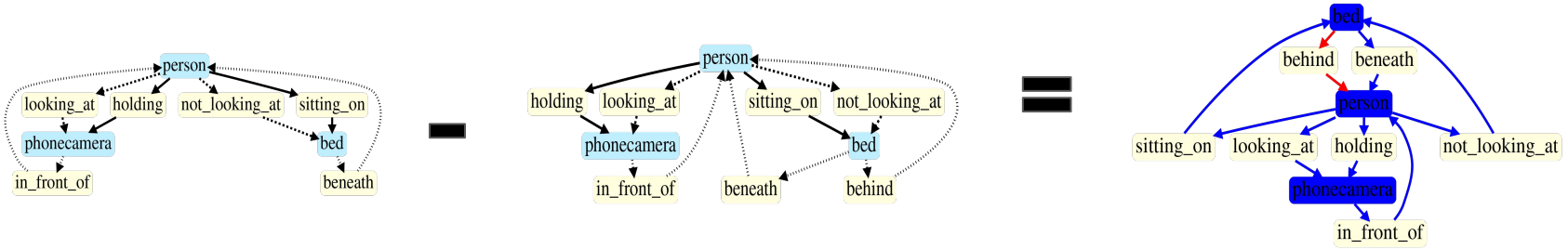
Removed relationships

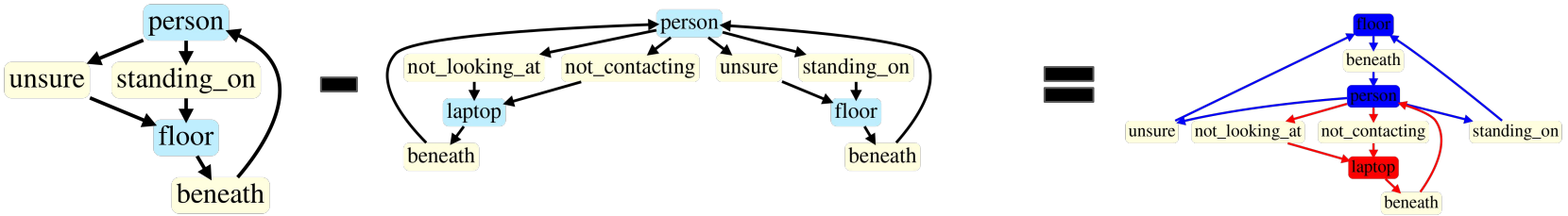
Added objects

Remaining objects

Removed objects

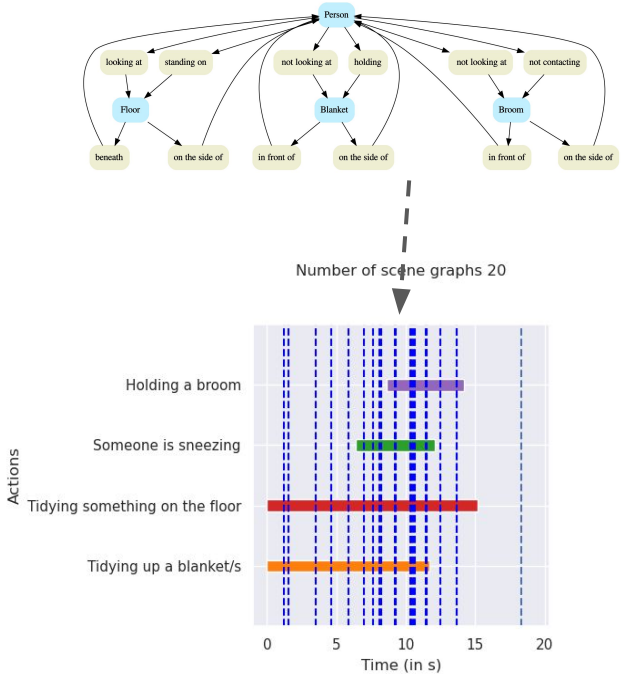
Predicates





Breaking up the GT Graph

- Working without Spatial relationships
  - Avoid noisy annotations
  - **Baseline** has all the relationships
- Can we know which part of scene graphs results from which action ?
  - How do you predict it ?

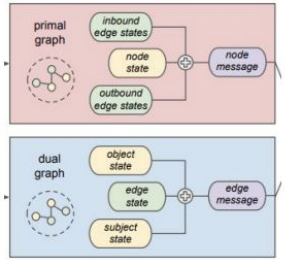


Spectrum of work in Image SG Generation

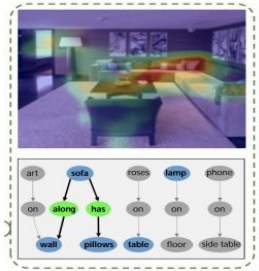
- Image-based SG Generation
  - Pixels to SG Generation
    - **[MSDN]** - Newell, Alejandro, and Jia Deng. "Pixels to graphs by associative embedding Neurips" 2017
  - Object proposals to SG Generation
    - **[IMP]** - Xu, Danfei, et al. "Scene graph generation by iterative message passing." CVPR 2017
    - **[Neural Motif]** - Zellers, Rowan, et al. "Neural motifs: Scene graph parsing with global context CVPR 2018
    - **[Graph-RCNN]** - Yang, Jianwei, et al. "Graph r-cnn for scene graph generation ECCV 2018
    - **[RelDN]** -Zhang, Ji, et al. "Graphical contrastive losses for scene graph parsing CVPR 2018
- Common Sense
  - Use wordnet along with dataset SG
    - Zareian, Alireza, Svebor Karaman, and Shih-Fu Chang. "Bridging knowledge graphs to generate scene graphs.", ECCV 2020
    - Zareian, Alireza, et al. "Learning Visual Commonsense for Robust Scene Graph Generation.", ECCV 2020
- Causality and SG
  - Tang, Kaihua, et al. "Unbiased scene graph generation from biased training.", CVPR 2020

Spectrum of Work

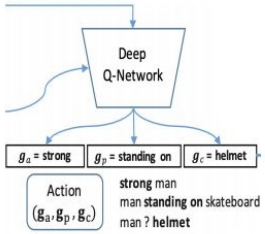
Message passing



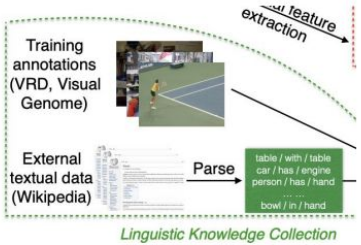
Attention



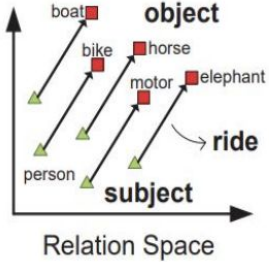
Reinforce



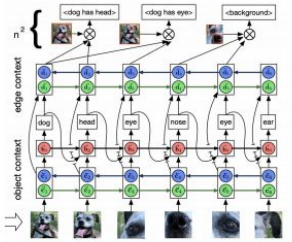
External knowledge



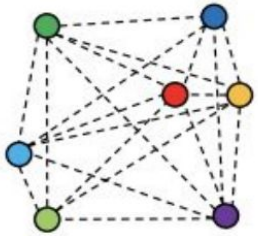
Transformations



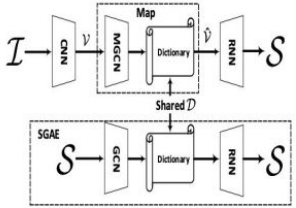
Recurrent networks



Graph Convolutions

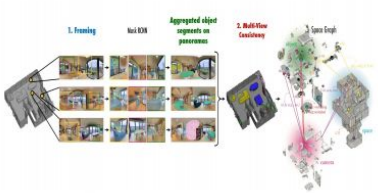


Auto-encoders

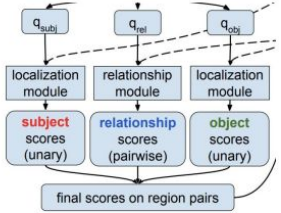


Spectrum of Work

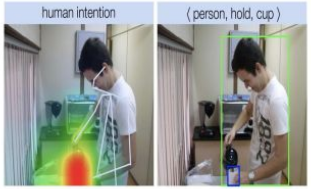
3D scene graphs



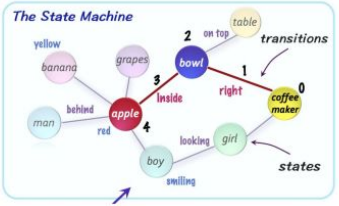
Explainable AI



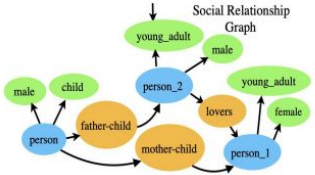
Human intentions



VQA

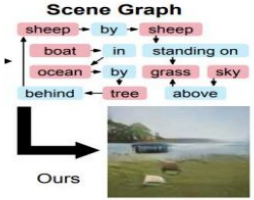


Social relationships

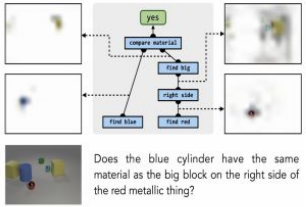


Fashion

Image generation



Program synthesis



Thank You