Lecture 3:
Semantic & Temporal Segmentation and Relation Grounding

Papers for Lecture 3: Semantic & Temporal Segmentation, & Relation Grounding

P3-1: Semantic Segmentation: (Presenter: Cheng Yi) (Asker: Wu Yihang)

(Must-Read) A Kirillov, E Mintun, N Ravi, et al. Segment anything. arXiv 2023.

(To-Read) K He, G Gkioxari, P Dollár & R Girshick (2017). Mask R-CNN. ICCV 2017.

P3-2: Temporal Segmentation: (Presenter: Thong Nguyen) (Asker: Dai Yuhe)

(Must-Read) Z Hou, W Zhong, L Ji, D Gao, K Yan, et al. CONE: An Efficient COarse-to-fiNE Alignment Framework for Long Video Temporal Grounding. ACL 2023.

(Must-Read) LA Hendricks, O Wang, E Shechtman, J Sivic, T Darrell & B Russell. Localizing Moments in Video with Temporal Language. EMNLP 2018.

P3-3: Relation Grounding: (Presenter: Zheng Jingnan) (Asker: Dibyadip Chatterjee)

(Must-Read) Y Cong, MY Yang & B Rosenhahn. RelTR: Relation Transformer for Scene Graph Generation. TPAMI. 2023.

(To-Read) B Dai, Y Zhang & D Lin. Detecting Visual Relationships with Deep Relational Networks. CVPR 2017

From Semantic to Temporal Segmentation

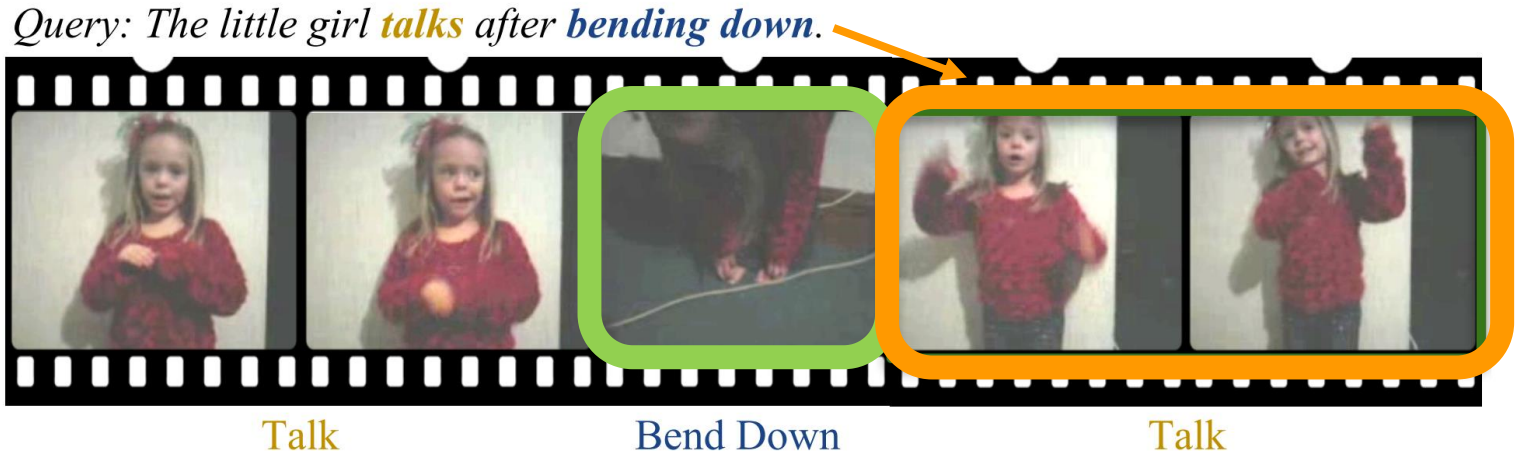
- **(Spatial) Segmentation of Anything:**

Prompt it with interactive points and boxes

Automatically segment everything in an image

Generate multiple valid masks for ambiguous prompts

- **Temporal Segmentation of any Concept:**



Perform temporal segmentation w.r.t.:

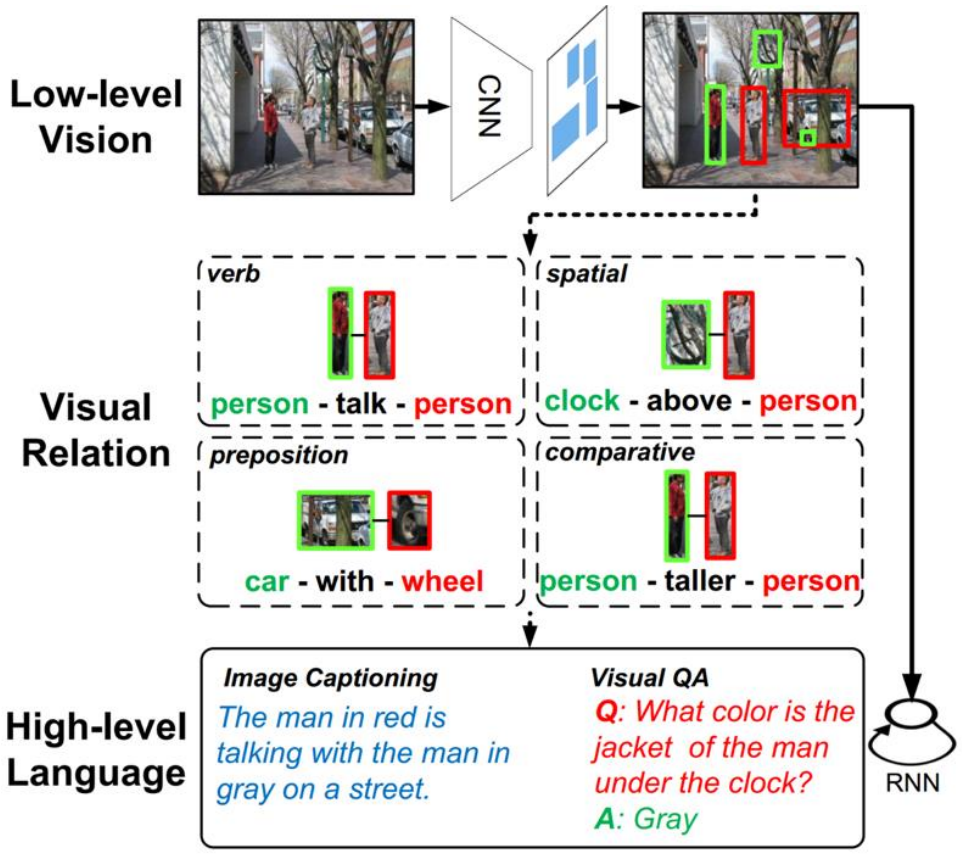
- Entities – simple
- Entities + Context
- Any arbitrary concepts

Key challenges:

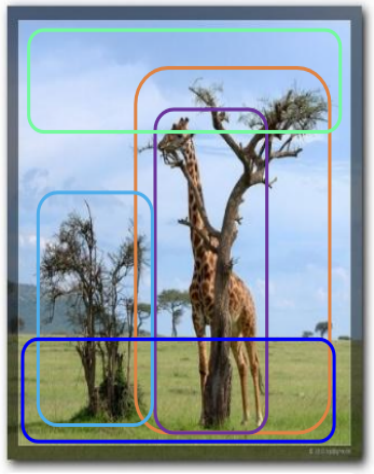
- To identify key entities from trivial ones
- This is query and context related

Beyond Spatial-Temporal Segmentation: Relation Inference

- **Relation inference:** formulated as the detection of relationships among entities
  - Permit us to find all (meaningful) triplets of {Sub, Pre, Obj} in images/videos
  - Bring the level of video semantics to that of text/ language
  - Support the integration of text and video in cross-modal analysis



- Again, need to focus on important relations from those trivial ones
  - The problem is query and context dependent
  - A very long-tailed problem for meaningful relations



(sky, above, tree)
(tree, under, sky)
(tree, on, grass)
(giraffe, on, grass)
(giraffe, behind, tree)
(giraffe, eat, tree)

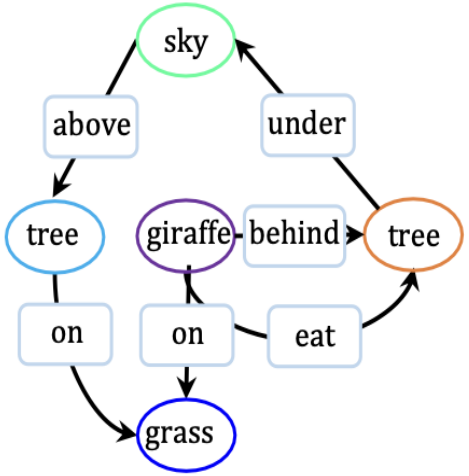
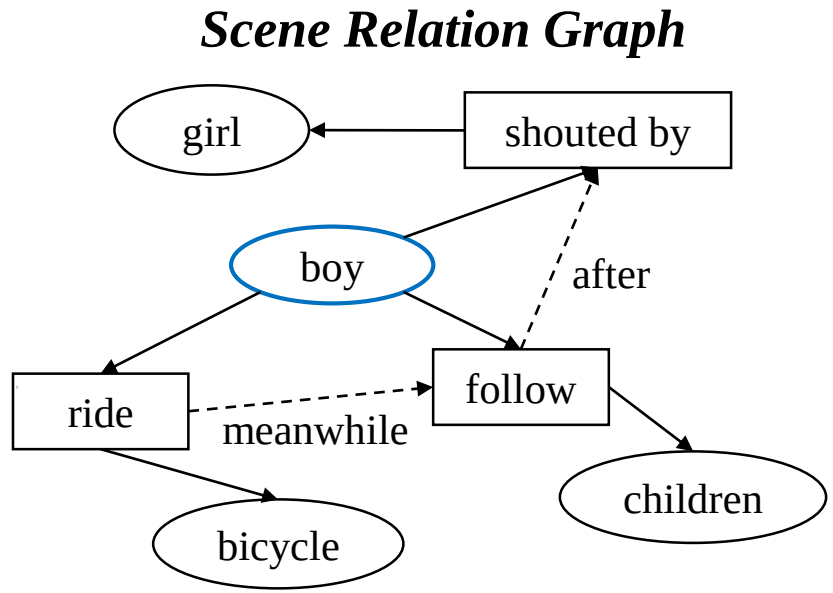


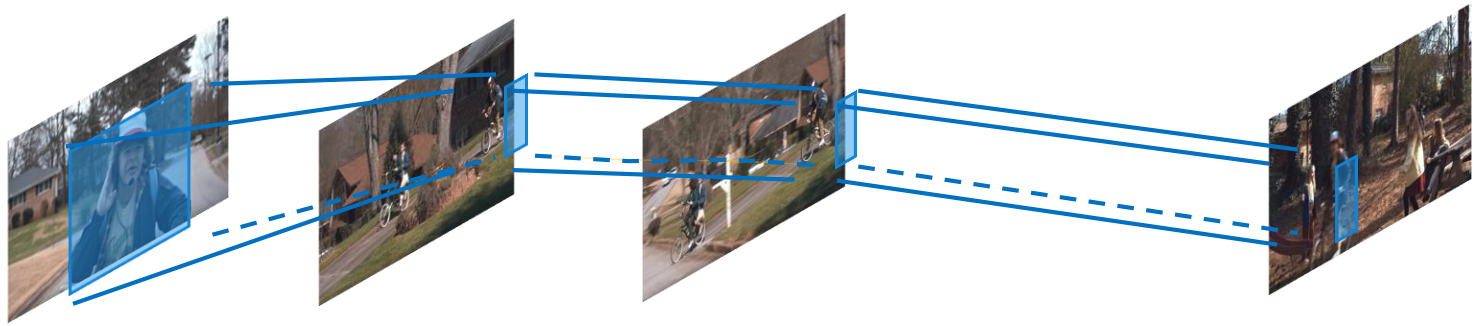
Image | Triplets | Graph

Spatial-Temporal Segmentation with Relation Inference

- Towards dynamic scene graph
  - A complete description of (essential activities) in the scene
  - May infer higher order relations and scene semantics
  - May be implicit or explicit
- Should be extended to multimodal scene graphs – towards visual-language integration



Spatio-Temporal Grounding



Requirements for Paper Presenters and Askers

- **Presenter:** The presentation of a sub-topic should cover (25 mins):
  - Objectives of papers
  - Clear literature reviews
  - Limitations, design/ implementation and results
  - Highlight key innovations, answer the how and why questions, such as How it works and Why it works
  - Future work.
  - Presenter Report: the presenter needs to submit a report within 2 weeks time (≤ 2 pages, Single-Spaced Times font 12)
- **Asker:**
  - You will need to pose 2-3 questions
  - Questions should have good depth and help to uncover insight of paper

Requirements for Short Idea/ Opinion Articles

- **Topics for Short Idea Paper:**
  1) Can LFM (Large Foundation Model) use public data for training and content generation: what are the issues and guidelines?
  2) Robust and trust are the key concerns to LFM. Is this a fundamental unsolved problem? What can be done to address this problem?
  3) Visual content has always been a supplementary feature to text in semantic analysis. Is this just a problem of maturity of visual analysis tools, or a fundamental problem in multimodal semantic understanding? How LFM should be designed to address this problem?
- **Requirements for the Paper:**
  - The writeup should cover the background, issues, positions, analysis and insights.
  - I am looking for new angles into the issues, as well as innovative ideas and insights.
  - I will award a **B** if paper covers most points above, and **A** for innovative ideas and insights
  - The article should be within 4 pages, in single-spaced Times Roman font 12 (excluding references).
- **Deadlines:**
  - Article 1: 16 Feb @1700.
  - Article 2: 8 Mar @1700.

Papers for Lecture 4: Cross-modal Alignment and Multimodal Scene Graph

P4-1: Cross-modal Alignment: (Presenter: Chai Zenghao)

(Must-Read) J Li, D Li, S Savarese & S Hoi. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. ICML 2023.

(Must-Read): A Radford, JW Kim, C Hallacy, et al. Learning Transferable Visual Models from Natural Language Supervision. ICML 2021.

(To-Read): L Qu, M Liu, J Wu, Z Gao & L Nie. Dynamic Modality Interaction Modeling for Image-Text Retrieval. SIGIR 2021.

P4-2: Multimodal Scene Graph: (Presenter: Dibyadip Chatterjee)

(Must-Read) J Yang, W Peng, X Li et al. Panoptic Video Scene Graph Generation. CVPR 2023.

(To-Read) K Tang, Y Niu, J Huang et al. Unbiased Scene Graph Generation From Biased Training. CVPR 2020.

(First Dataset, Must-Read) R Krishna, Y Zhu, O Groth, et al. Visual Genome: Connecting language and vision using crowdsourced dense image annotations. IJCV 2017