# Semantic Understanding with Interaction Tracking in Long-form Multimodal Sources

Vishal Anand
vishal.anand@columbia.edu
vishal.anand@microsoft.com

Columbia University, Microsoft

COLUMBIA UNIVERSITY
IN THE CITY OF NEW YORK

NIST
National Institute of Standards and Technology
U.S. Department of Commerce

TRECVID, 2022

# Research Goal

Learn semantic relationships over evolving long form multi-modal data

# Previous Works

## MultiModal Language Modelling on Knowledge Graphs for Deep Video Understanding

ACM Multimedia 2021 Paper | ACM Multimedia 2020 Paper | ICMI Workshop 2020 Paper (Oral) | Interactive Interface


ACM Multimedia 2021:
Vishal Anand[1,3] Raksha Ramesh[1,2] Boshen Jin[1,2] Ziyin Wang[1,2] Xiaoxiao Lei[2] Ching-Yung Lin[1,2]

ICMI DVU 2020:
Raksha Ramesh[1,2] Vishal Anand[1,3] Ziyin Wang[1,2] Tianle Zhu[1] Wenfeng Lyu[1] Serena Yuan[1] Ching-Yung Lin[1,2]

ACM Multimedia 2020:
Vishal Anand[1,2] Raksha Ramesh[1] Ziyin Wang[1,2] Yijing Feng[1] Jiana Feng[1] Wenfeng Lyu[1] Tianle Zhu[1] Serena Yuan[1] Ching-Yung Lin[1,2]

[1]Columbia University, New York, NY, USA
[2]Graphen AI, New York, NY, USA
[3]Microsoft, Redmond, WA, USA

# Works - 2022

- Vishal Anand, Yifei Dong, Raksha Ramesh, Zifan Chen, Yun Chen, Linquan Li, Ching-Yung Lin. 2022. Semantic Understanding and Evolving Interaction Tracking in Long-form Multimodal Datasets. In Proceedings of TRECVid 2022.

- Vishal Anand, Raksha Ramesh, Boshen Jin, Ziyin Wang, Xiaoxiao Lei, Ching-Yung Lin. 2022. Leveraging Text Representation and Face-head Tracking for Long-form Multimodal Semantic Relation Understanding. In Proceedings of the 30th ACM International Conference on Multimedia (MM '22). https://dl.acm.org/doi/abs/10.1145/3503161.3551610

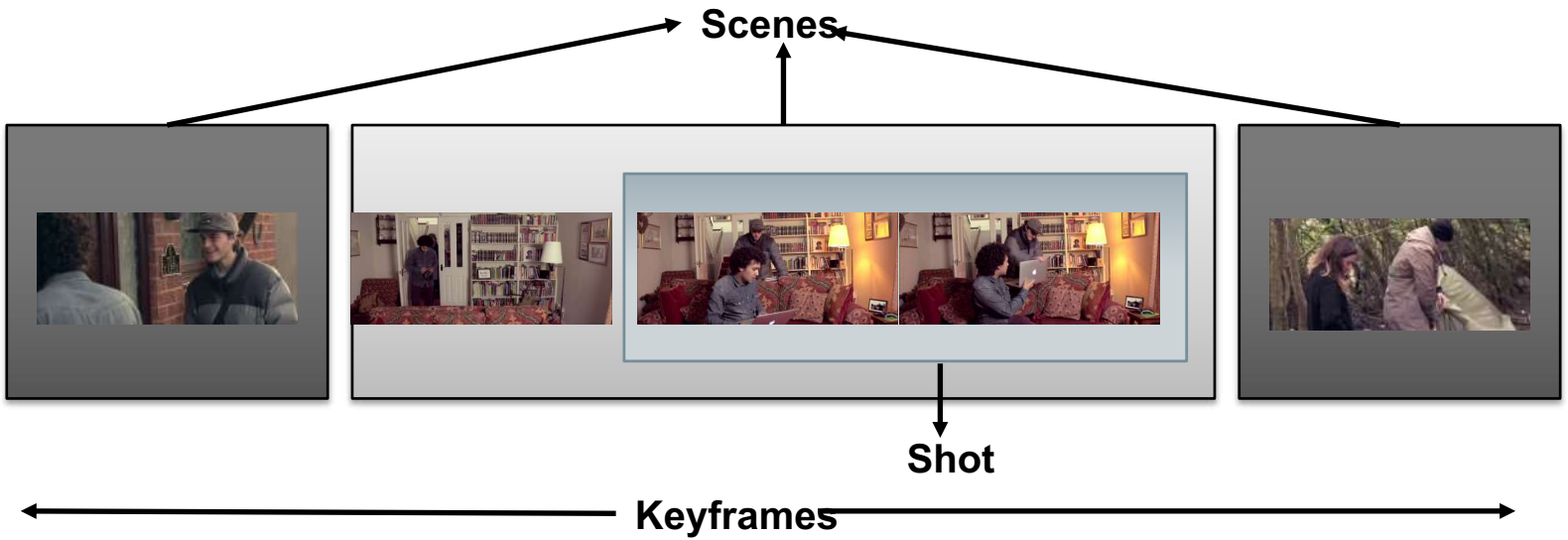
# Works – 2020, 2021

- Vishal Anand, Raksha Ramesh, Boshen Jin, Ziyin Wang, Xiaoxiao Lei, Ching-Yung Lin. 2021. MultiModal Language Modelling on Knowledge Graphs for Deep Video Understanding. In Proceedings of the 29th ACM International Conference on Multimedia (MM ’21). https://doi.org/10.1145/3474085.3479220

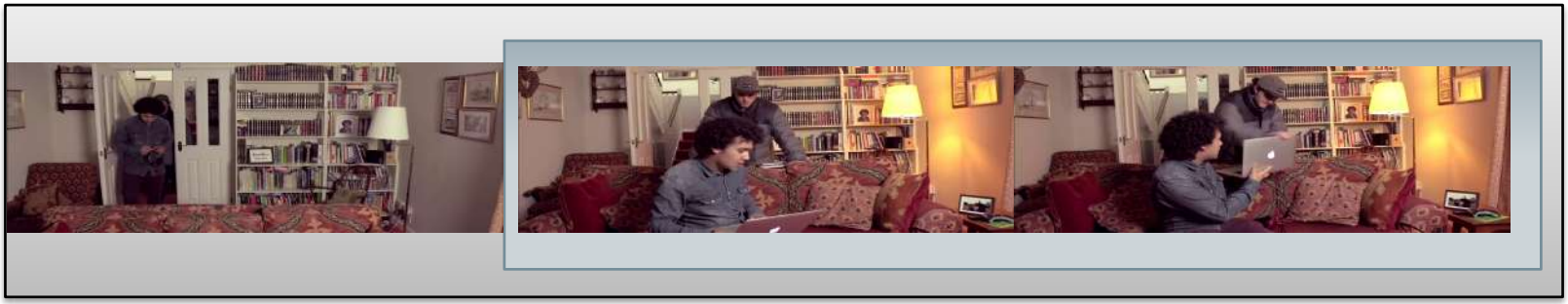
- Raksha Ramesh, Vishal Anand, Ziyin Wang, Tianle Zhu, Wenfeng Lyu, Serena Yuan, and Ching-Yung Lin. 2020. Kinetics and Scene Features for Intent Detection. In Companion Publication of the 2020 International Conference on Multimodal Interaction (Virtual Event, Netherlands) (ICMI '20 Companion). Association for Computing Machinery, New York, NY, USA, 135--139. https://doi.org/10.1145/3395035.3425641

- Vishal Anand, Raksha Ramesh, Ziyin Wang, Yijing Feng, Jiana Feng, Wenfeng Lyu, Tianle Zhu, Serena Yuan, and Ching-Yung Lin. 2020. Story Semantic Relationships from Multimodal Cognitions. In Proceedings of the 28th ACM International Conference on Multimedia (MM ’20). https://doi.org/10.1145/3394171.3416305

# Segments of Long-Form Videos



# Objective



Some questions we are interested in:
Who is Person A, Person B ?
What is Person A's interaction with Person B in this scene?
What is Person A's relationship with Person B?
What is the overall sentiment of the scene?
Where is Person A & Person B?

# Dataset

| | Statistics | | | |
|---|---|---|---|---|
| | #Actor | #Speaker | #Object | Time |
| Honey | 10 | 10 | 12 | 86 min |
| Nuclear Family | 4 | 4 | 5 | 28 min |
| Spiritual Contact | 10 | 10 | 13 | 66 min |
| Super Hero | 7 | 7 | 12 | 18 min |
| Huckleberry Finn | 10 | 10 | 20 | 106 min |
| Valkaama | 7 | 7 | 13 | 93 min |
| Shooters | 8 | 8 | 11 | 41 min |
| Let's Bring Back Sophie | 13 | 13 | 22 | 50 min |
| The Big Something | 9 | 9 | 12 | 101 min |
| Time Expired | 16 | 16 | 36 | 92 min |

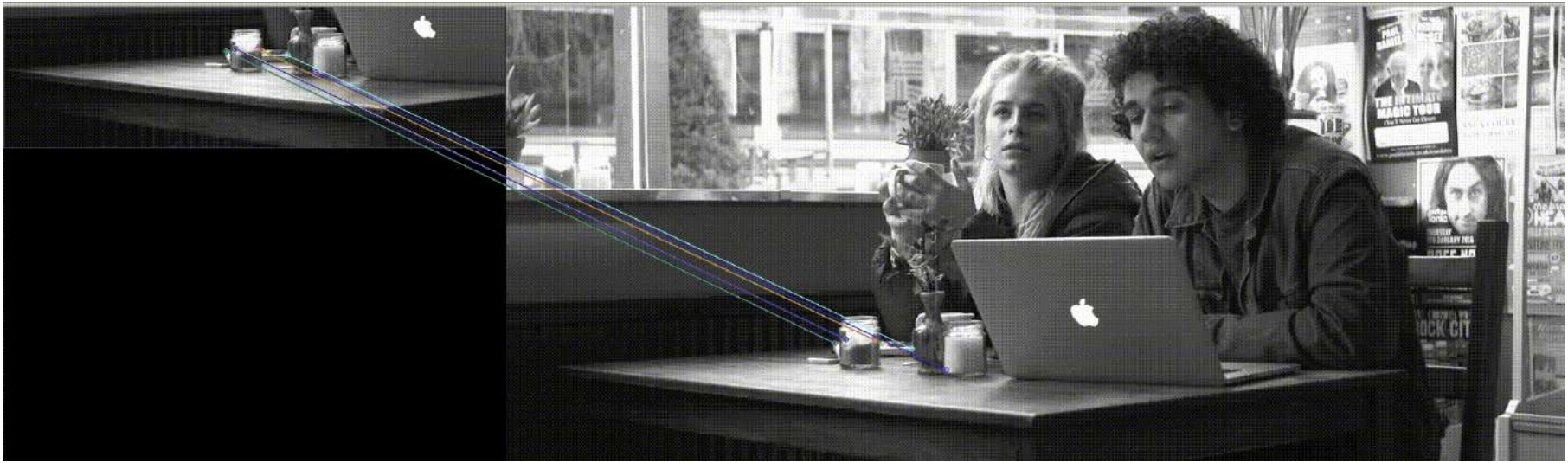
- Dataset annotated with entities: persons, objects, locations
- Pairwise relations between entities
- Actions & events

# Speaker Kinetics

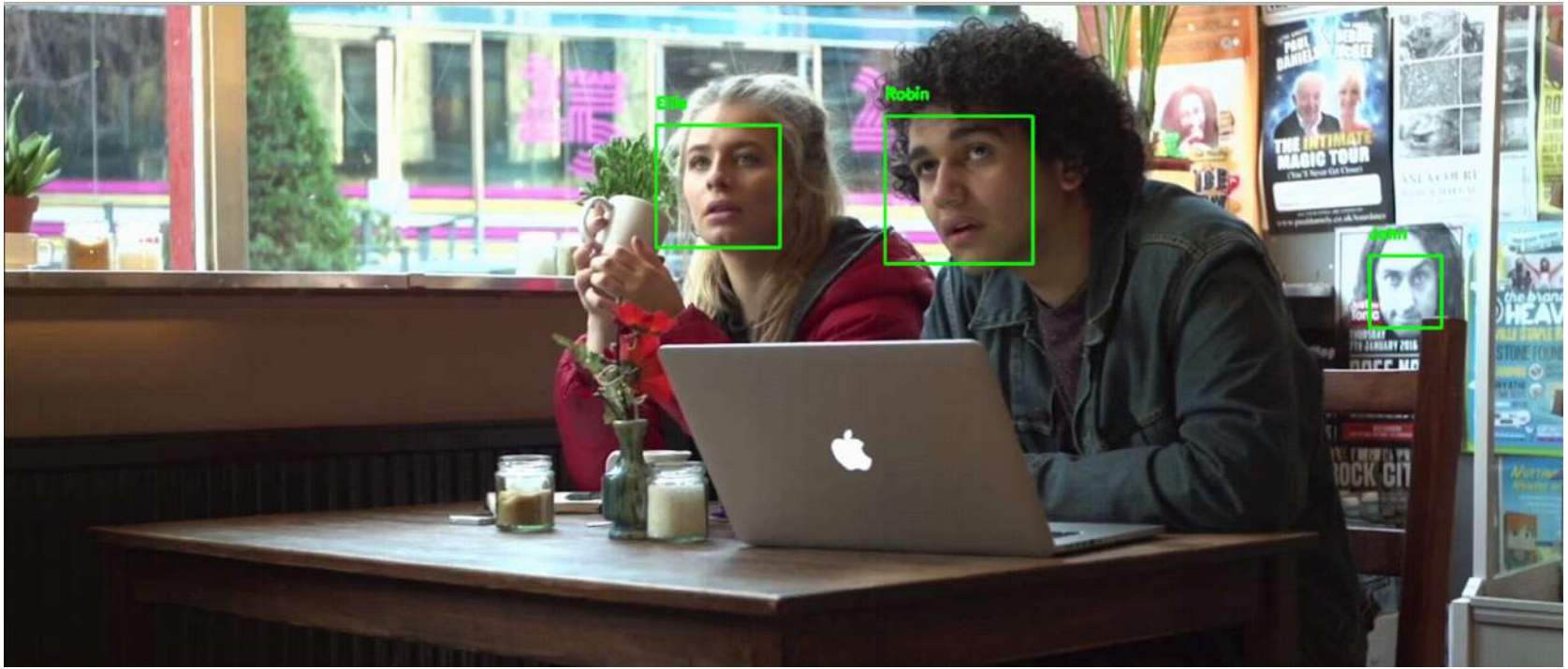
# Speaker Kinetics

# Speaker Kinetics

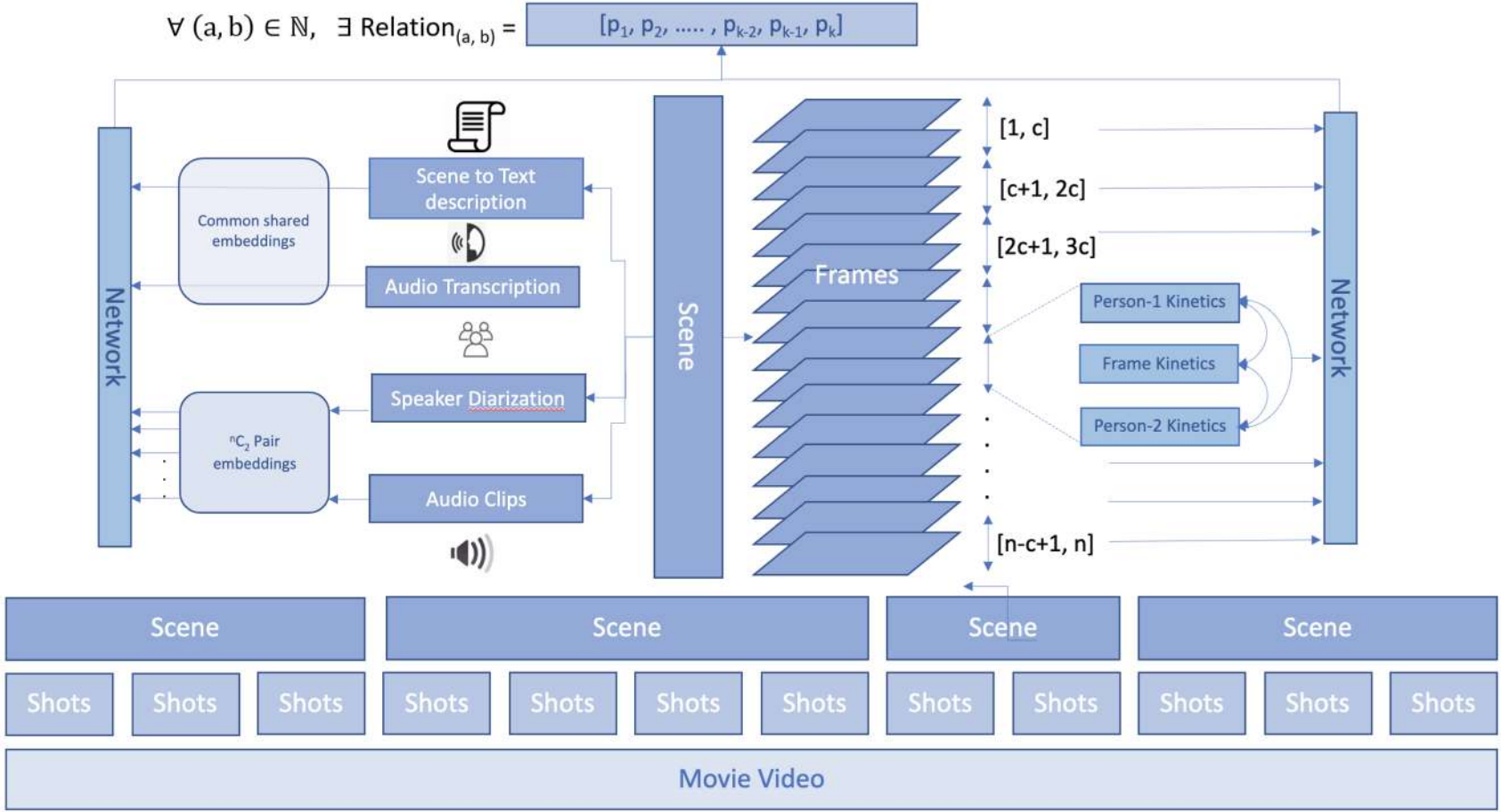
# Object Entity Mapping



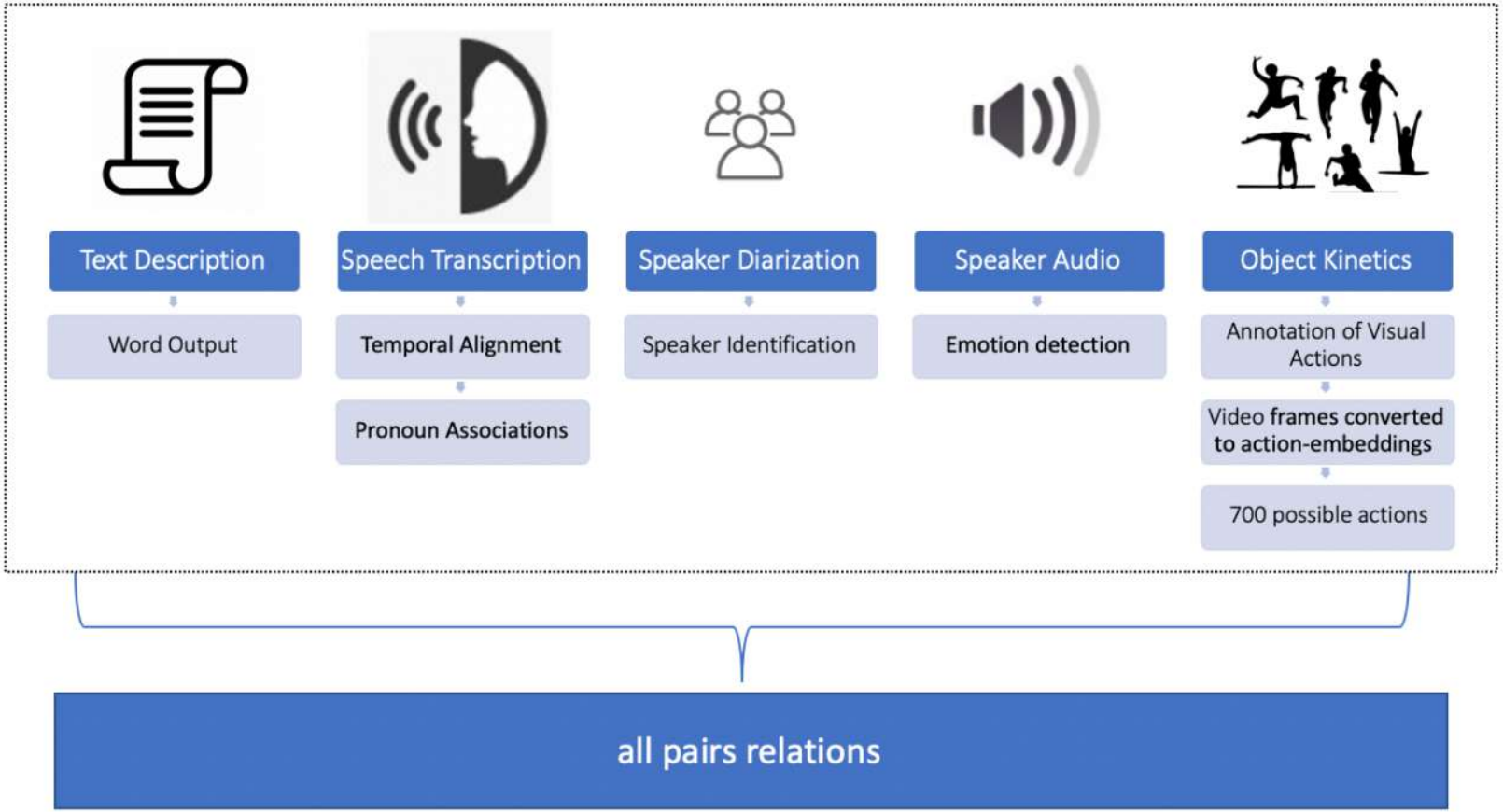
# Face Detection



# Initial Architecture

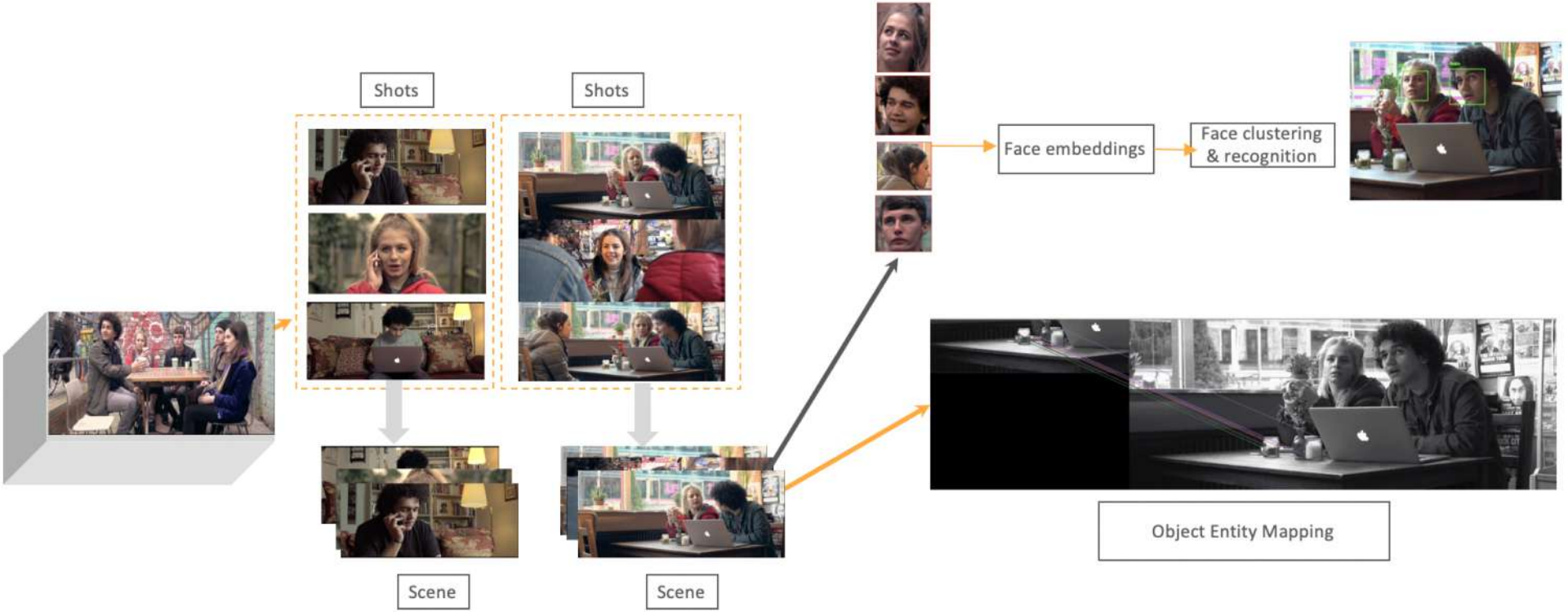


# Modalities

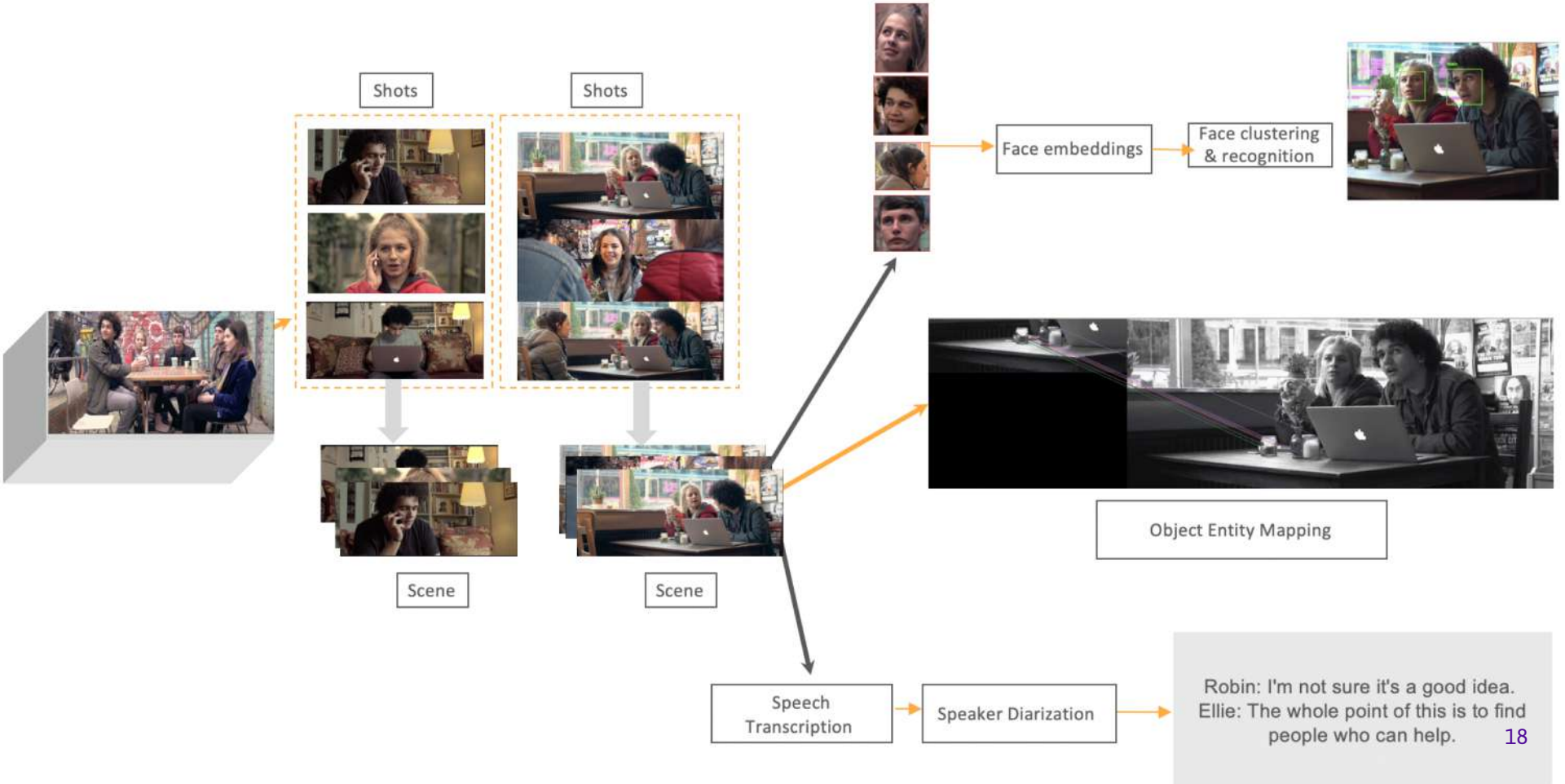

Text Description
Word Output
Speech Transcription
Temporal Alignment
Pronoun Associations
Speaker Diarization
Speaker Identification
Speaker Audio
Emotion detection
Object Kinetics
Annotation of Visual Actions
Video frames converted to action-embeddings
700 possible actions
all pairs relations

# Entity Localization

# Entity Localization



# Entity Localization

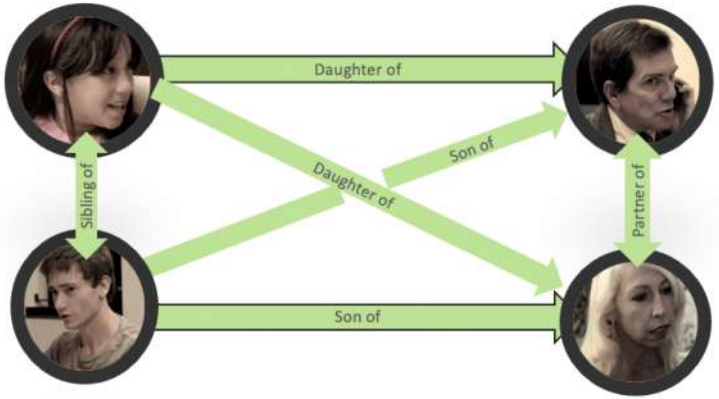


# Semantic Relationships

# Semantic Relationships

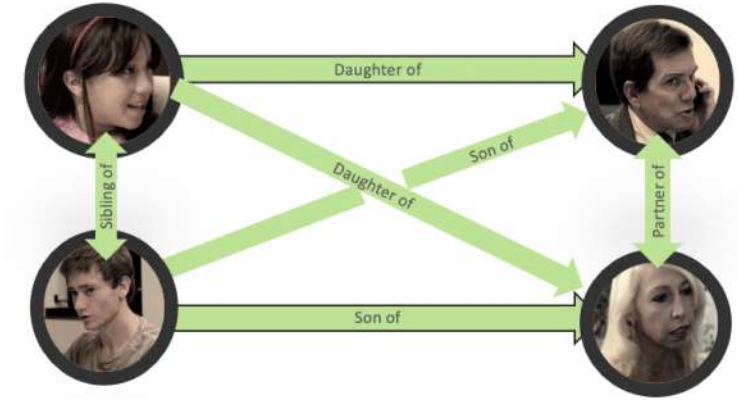
# Semantic Relationships

Visual cues & Context



# Semantic Relationships


Visual cues & Context




P: Nan, I've brought Jeremiah today because he needs help.

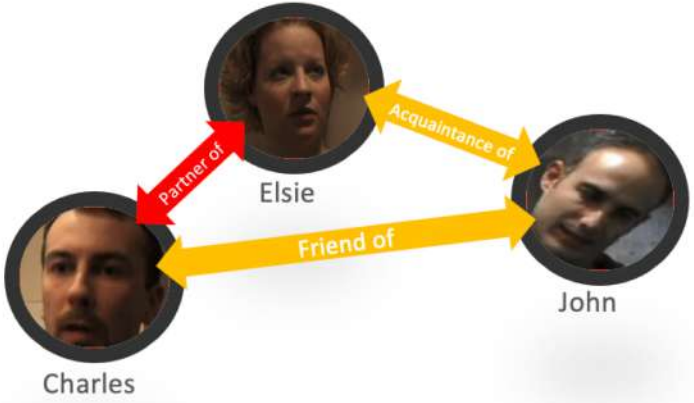
G: Okay how can I help you Jeremiah?

J: I've been attacked by spirits

G: When Paul's grandad died, I went through something similar.

# Semantic Relationships



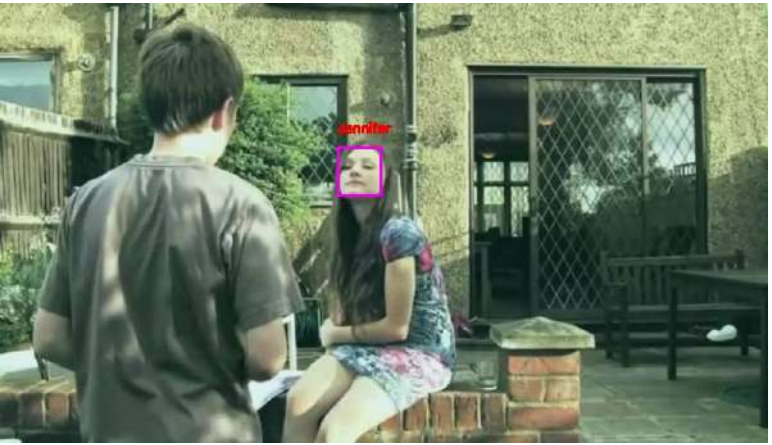



# Multi-Body Tracking

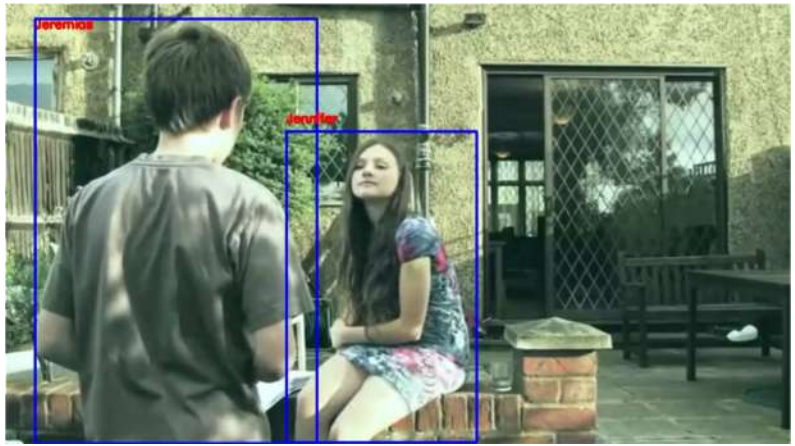
Who is Person A, Person B ?

1. Free-form movies capture subjects from all angles leads to information loss with conventional face-detectors

2. Fine-grained interaction predictions require accurate face recognition & tracking



(a) Face recognition only



(b) Face recognition with multi-body tracking

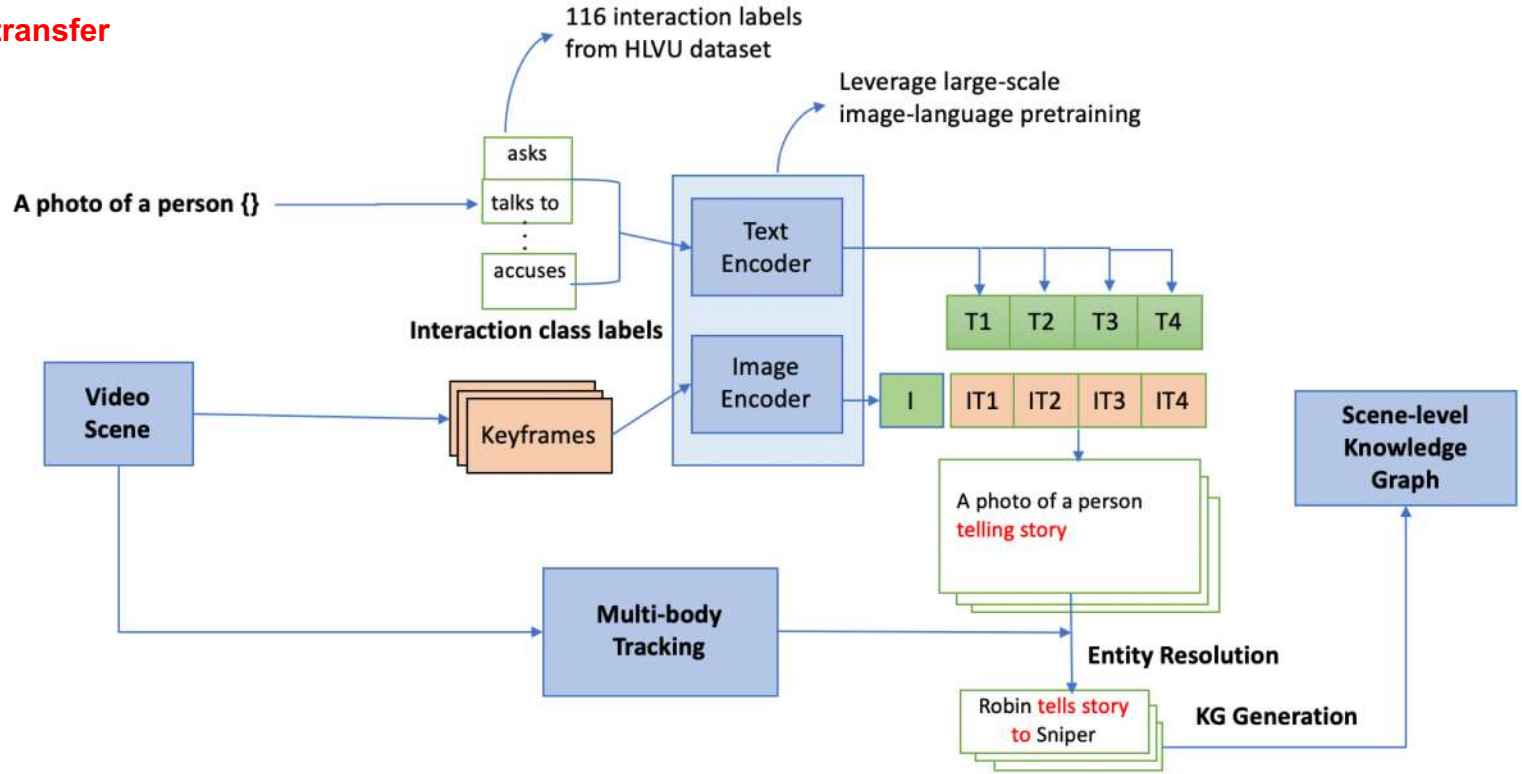
Tracking by overlap between the Mask R-CNN “person” object mask and face recognition bounding box.

As long as face is recognized once in a shot they can be tracked throughout the shot

# Framework

What is Person A's interaction with Person B in this scene?

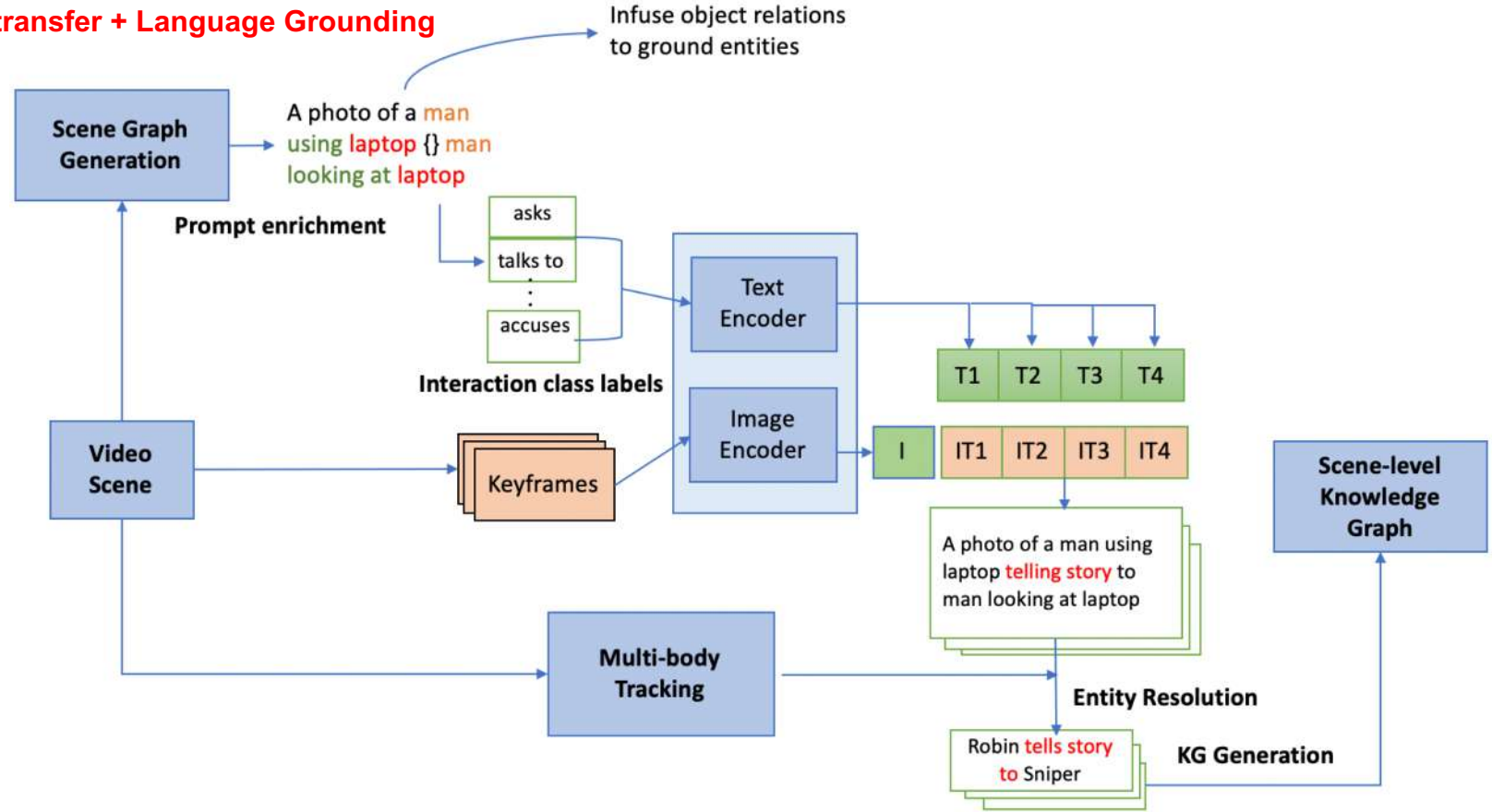
**Zero Shot transfer**



# Framework

What is Person A's interaction with Person B in this scene?

**Zero Shot transfer + Language Grounding**



# Key frame extraction

## Prior method:

**i-frame**
Disadvantage: capture a number of **unrelated frames**

## Current approach:

**Multi-entity-frame**

Per head tracking results, capture frames with **2+ entity_faces**, and **sleep 21 frames** between each capture

Figure 1: Bagman-5-3670

Figure 2: Bagman-5-3672

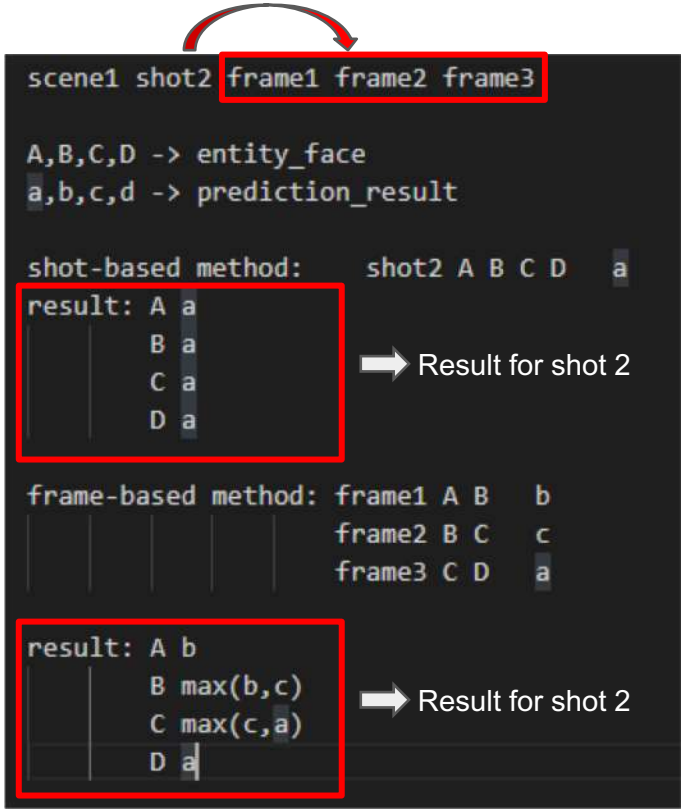
# Key frame extraction

## Old method (shot-based):

Map all prediction results in shot A with all face info in shot A (including +1 and -1 shots)

## New method (frame-based):

Map each prediction results of frame A with every face information in frame A



# Challenges: multi-entity-frame process

## 1. Loss of relevant information

From the neighbor frame of movie bagman

Can be captured by multi-entity-frame extraction and i-frame extraction

Only can be captured by i-frame extraction

**Solution**: we merge the face recognition results of a range of frames(+/- 5) with the original face recognition results of that key frame

# Challenges: multi-entity-frame process

2. Not applicable for scenes with strictly one-entity head tracking result

```
Ruth,002,323,85,613,327,48,18,1600.0
Ruth,002,321,88,614,328,49,18,1633.3333333333335
Ruth,002,320,89,614,328,50,18,1666.6666666666667
Ruth,002,318,89,615,329,51,18,1700.0000000000002
Ruth,002,318,88,611,328,52,18,1733.3333333333335
Ruth,002,317,87,610,328,53,18,1766.6666666666667
Ruth,002,317,88,610,327,54,18,1800.0000000000002
```

Head tracking results from movie honey scene 18

**Solution**: Merge face recognition results over frame-range (+/- 5) with original face recognition results of key frame

# Challenges: Head Tracking

## 2. Missing head tracking results for some scenes generates no clip predictions

**Solution**: Using face recognition results from our ACM MM 2021 paper for cross-referencing

Four sets of experiments:

1. Only use multi-entity tracking
2. Merge multi-entity tracking and face recognition with the same priority
3. Merge multi-entity tracking and face recognition with multi-entity tracking prioritized
4. Merge multi-entity tracking and face recognition with face recognition prioritized

# Evaluation

## Question Generation

Step 1 - generate knowledge graph using ground truth

E.g. Ground truth :

(...), ..., (4, Runa), (5, Max), (6, Ari) ..., (...) # (...), ..., (5, 4, Sibling Of), (5, 6, Sibling Of) ..., (...)

The graph we get from it (partial):

Graph = {..., Max: [..., (Sibling Of, Runa), (Sibling Of, Ari), ...], ...}

Step 2 - generate question queries according to the KG generated above

E.g. Q: How many siblings does Max have? Choices: 1,0,4,6,2

A: 2

## Question Answering & Evaluation

Using KG generated by model's prediction result to answer self generated questions.

Using Mean Reciprocal Rank(MRR) and Accuracy to evaluate model's prediction performance

# Evaluation result analysis

## Scene level questions

S1 - Polarization pattern

The first question asks us to find specific scene according to a set of interactions, and the result (average MRR score for each movie) shows a pattern with polarization.

Movies with relatively **LOW** average MRR score for this question:

"Calloused Hands" -> (0.072), "Chained For Life" -> (0.016), "Liberty Kid" -> (0.046), "Losing Ground" -> (0.125)

Movies with relatively **HIGH** average MRR score for this question:

"Like Me" -> (0.539), "Little Rock" -> (0.289)

S2 & S3 - No meaningful pattern

These scene level questions that ask the previous/next interaction after one specific interaction in a scene, we could not get meaningful pattern from the result. The reason for this is our model does **NOT** involve any **temporal factor** in both input and output, and we basically "guess" the answer

# Evaluation result analysis

## Movie level questions

S1 - The possibility of answer could be location entity is neglected

We did not consider the probability of answer to this question could be location entity as well. We only used person entity as our possible answer.

This problem could be improved by choosing the correct answer set (person or location) before answering the question according to the type of the relation(person-person or person-location) and subject type (person or location) in the question prompt.

S2 - Lack the process of omitting incorrect category of relation

Sometimes the question asks what the relation is between 2 people. We may give an answer that is person-location relation instead of person-person relation because the knowledge graph does not contain one relation between these 2 people that is in the question's choices. Then we will choose the closest relation in choice according to the relation similarity matrix if these 2 people do have a relation, otherwise we will give a random choice.

We could improve this by filtering the choice first to omit the incorrect type of relation(i.e., person-person or person-location) according to the prompt of the question.

# Language model

| | |
|---|---|
| Prior method: | image encoder extracts features from key frames in each scene. |
| Improvement: | explicit language model for scene subtitles. |
| Baseline: | CNN, LSTM, BiLSTM, BERT |
| Dataset: | dialog_re for baseline training, trecvid dataset for fine tuning |
| Dialog Retrieval: | entities, type of entities, list of relations, list of relation trigger words. |

# Language model

## Movie subtitles should be converted into following format:

```
{
    'dialog': ["Speaker 1: <line 1>”, “Speaker 2: <line 2>, … ],
    'relational_data’: {
           'r': [['per:alternate_names'], ['per:alumni'], …  ],
           'rid': [[30], [4], …],
           't': [[''], [''], …],
           'x': ['Speaker 2', 'Speaker 4', …],
           'x_type': ['PER', 'PER', …],
           'y': ['Speaker 4', 'Tommy', …],
           'y_type': ['PER', 'PER', …]
    }
}
```

- `dialog`
  - List of dialog spoken between the speakers
- List of annotations per dialog per argument
  - `x` : First entity
  - `y` : Second entity
  - `x_type` : Type of the first entity
  - `y_type`: Type of the second entity
  - `r` : List of relations
  - `rid`: List of relation IDs
  - `t`: List of relation Trigger words

# Prompt variation

| Movie | M1-MRR-U | M1-MRR-L | M2-A |
|---|---|---|---|
| Manos | 49.3 | 20.2 | 23.5 |
| Road_to_bali | 37.8 | 18.9 | 18.0 |
| Bagman | 23.1 | 15.6 | 19.8 |
| Honey | 64.8 | 29.3 | 29.3 |
| Shooters | 58.3 | 34.0 | 12.1 |
| Huckleberry_Finn | 42.4 | 25.9 | 13.3 |
| Sophie | 43.6 | 24.5 | 22.4 |
| Spiritual_Contact | 40.7 | 29.3 | 19.6 |
| Valkaama | 51.0 | 37.0 | 30.0 |
| Nuclear_Family | 100 | 52.1 | 9.5 |
| Superhero | 78.6 | 37.0 | 34.2 |
| Average | 53.6 | 29.4 | **21.1** |

Table 9: Training Evaluation (percentage) on movie level tasks without location prompt

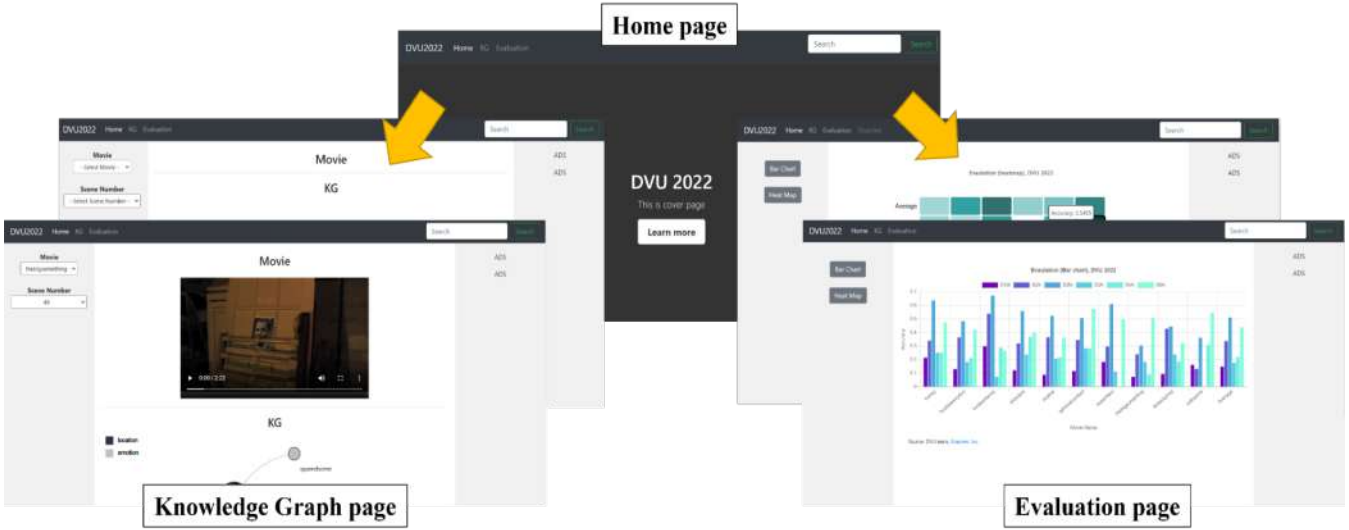
| Movie | M1-MRR-U | M1-MRR-L | M2-A |
|---|---|---|---|
| Manos | 44.2 | 20.2 | 20.4 |
| Road_to_bali | 35.1 | 18.9 | 17.3 |
| Bagman | 26.0 | 15.6 | 13.0 |
| Honey | 69.0 | 29.2 | 29.3 |
| Shooters | 60.8 | 34.3 | 12.1 |
| Huckleberry_Finn | 56.9 | 25.9 | 8.3 |
| Sophie | 36.5 | 24.5 | 20.0 |
| Spiritual_Contact | 41.0 | 29.3 | 21.7 |
| Valkaama | 52.9 | 37.0 | 33.3 |
| Nuclear_Family | 100 | 52.1 | 14.3 |
| Superhero | 78.6 | 37.0 | 23.7 |
| Average | **54.6** | 29.4 | 19.4 |

Table 10: Training Evaluation (percentage) on movie level tasks with location prompt

To have a better performance on scene description. Locations are included in each prompt as the input for the text encoder. This idea comes from attaching locations in every prompt as common sense for a higher confidence score.

# Visualization

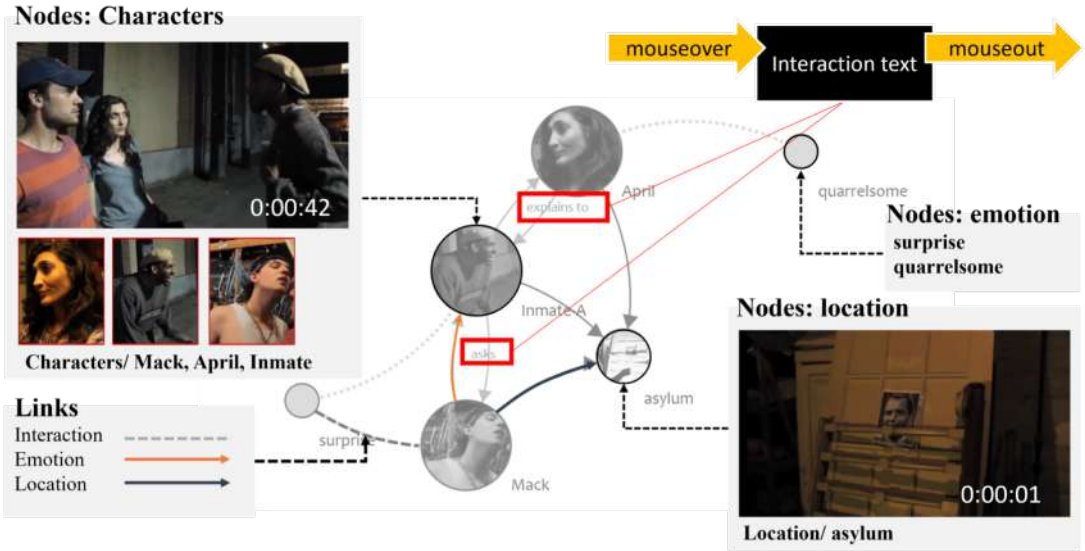
1. One-stop shopping website (OSS)
2. Strengthen the relevance of data
3. Increase the user's understanding of the experiment
4. Reflect the results of learning in a timely manner (timestamp)
5. Data visualization (scene by scene)



# Visualization - Knowledge graph

- Face tracking of characters
- Nodes: object name, node's type, character's face tracking
- Edges: relation type, relation, source and target

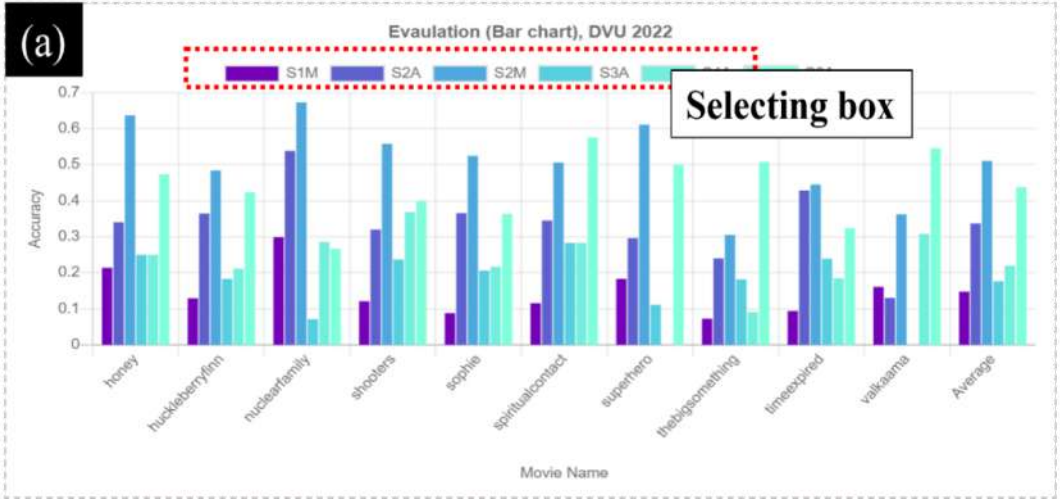
1. Location
2. Emotion
3. Interaction

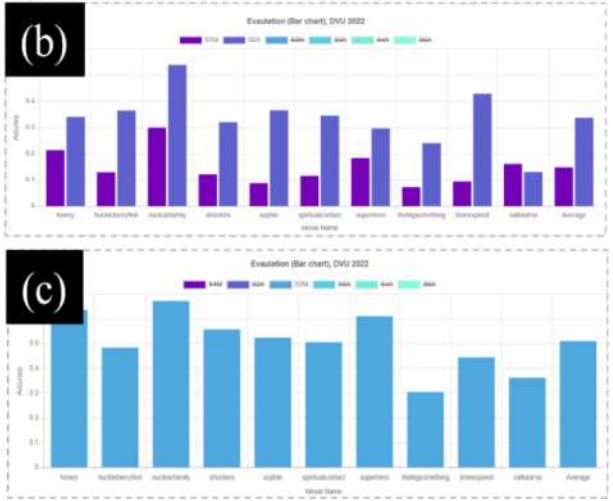


# Visualization - Evaluation

## Query and Answering visualization

1. Selecting box
2. Numerical and graphical displays
3. Compare multiple movie results



# Thank You

Vishal Anand
vishal.anand@columbia.edu
vishal.anand@microsoft.com

Columbia University, Microsoft

TRECVID, 2022