ACM Multimedia 2022
Lisbon, Portugal | 10-14 October

Multimodal Analysis for Deep Video Understanding with Video Language Transformer

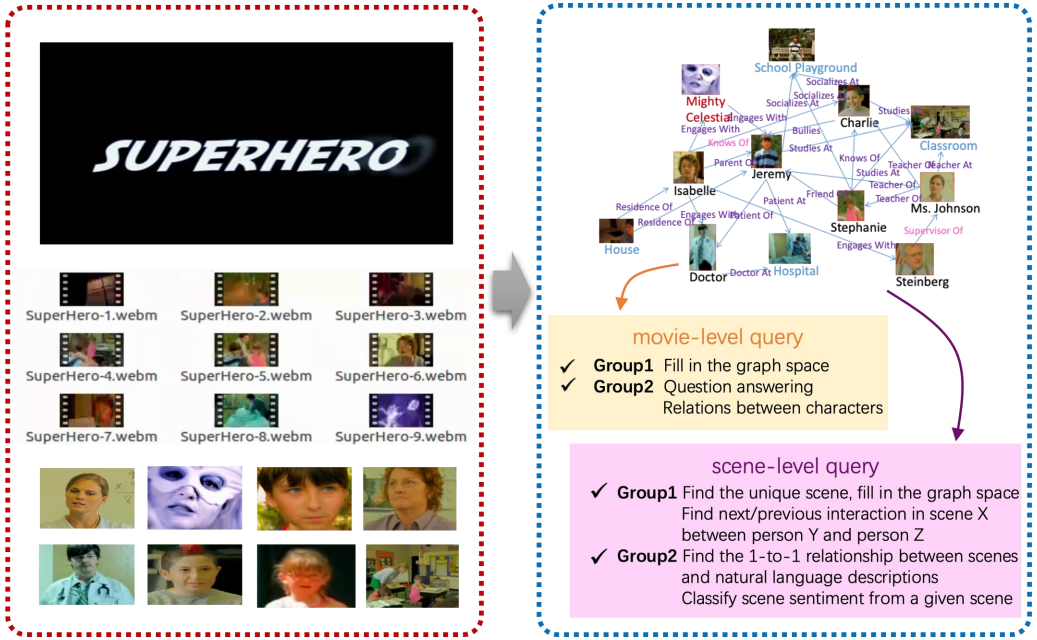
Beibei Zhang[1], Yaqun Fang[1], Tongwei Ren[1,*], Gangshan Wu[1]

[1]State Key Laboratory for Novel Software Technology, Nanjing University

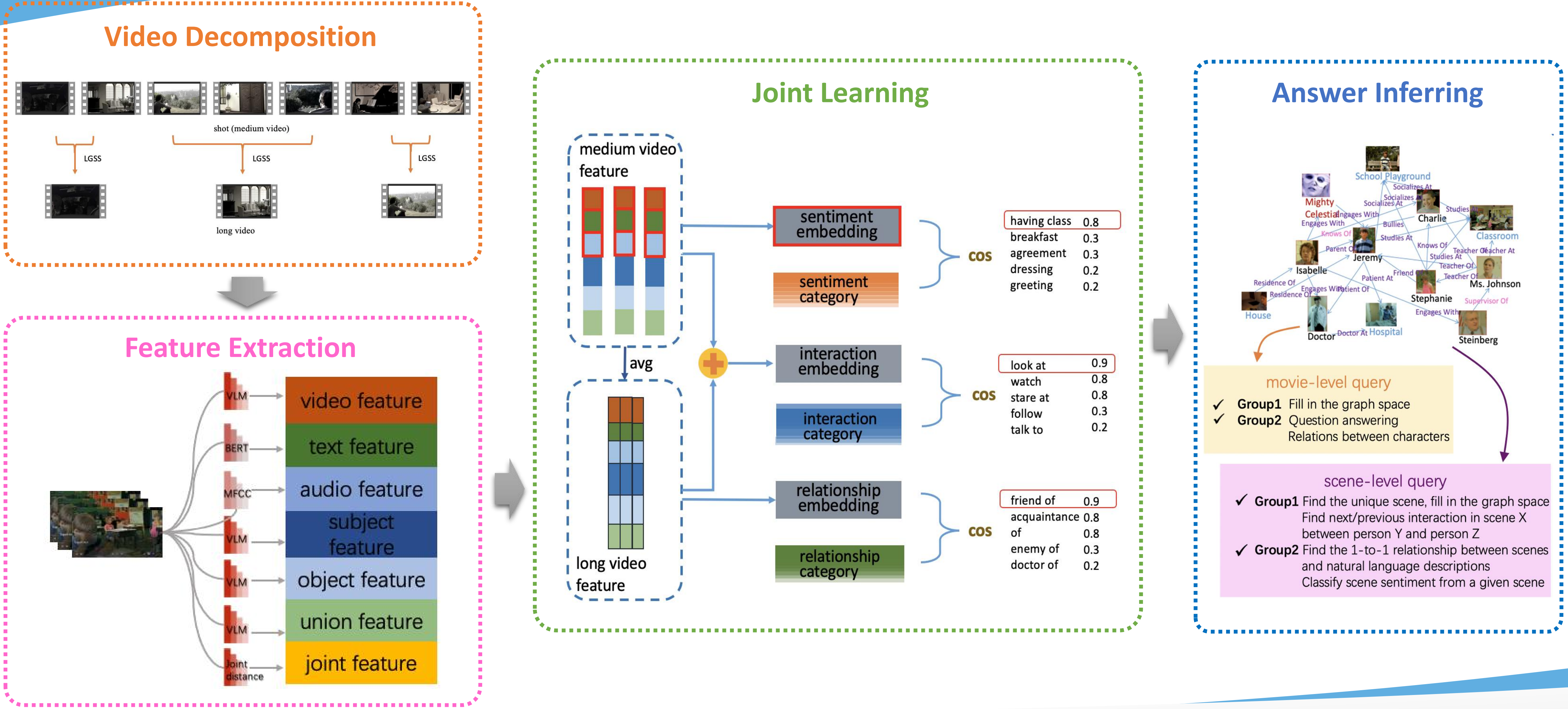
MAGUS
MediA recoGnition and UnderStanding

Task

- **Deep Video Understanding Challenge (2020 ~ 2022)**
  - **Movie-level**
    - **Group 1**
      - Find all possible paths question.
    - **Group 2**
      - Fill in the part of graph question.
      - Multiple choice questions.
  - **Scene-level**
    - **Group 1**
      - Find the unique scene.
      - Fill in the graph space.
      - Find next interaction in scene X between person Y and person Z.
      - Find previous interaction in scene X between person Y and person Z.
    - **Group 2**
      - Find the 1-to-1 relationship between scenes and natural language descriptions.
      - Classify scene sentiment from a given scene.



Our Pipeline



Answer Inferring: Movie-level

- **Fill in the graph space**
  - Sort the candidates in the entity-relationship graph according to scores generated by our method
- **Question answering**
  - Plug each choice into question and check whether the graph is satisfied
  - If none of the choices can fit our graph, choose a reasonable answer based on the types of entities and relationships
- **Relations between characters**
  - Collect the paths between two entities by depth-first searching through the graph

Answer Inferring: Scene-level

- **Find the unique scene, fill in the graph space**
  - Interaction knowledge graph
- **Find next/previous interaction in scene X between person Y and person Z**
  - Split medium video into shot videos
  - Predict the interaction of each shot video
  - **The prior knowledge is added to assist in judging the sequence of interactions**
- **Find the 1-to-1 relationship between scenes and natural language descriptions**
  - Match with predicted interactions and sentiments
  - **Match descriptions from the scene one by one**
- **Classify scene sentiment from a given scene**
  - Sentiment model prediction result
  - **Match video with sentiment directly by VLM**

Experiments

• Interaction prediction

- Metric: recall@10
- Component analysis: difference cases using different features and feature combinations for interaction prediction.
- Analysis:
  - **The combination of text feature and joint feature performs best**
  - **Difference among features is large.**

	All	Bagman	Manos	RoadToBali	TheIllusionist
V	4.70	3.40	12.70	3.10	2.60
V+P	0.90	1.10	2.60	0.00	0.40
V+T	10.90	**9.80**	**28.0**	10.90	**4.40**
V+T+A	7.00	2.90	16.90	8.00	3.10
V+T+A+P	2.30	1.40	10.10	0.30	0.90
T	8.50	3.70	24.30	8.50	2.60
T+P	**11.30**	6.00	27.00	**12.40**	**4.40**

V: visual feature; T: text feature; A: audio feature; P: joint feature of poses

Experiments

- **Sentiment query answering**
  - Metric: accuracy
  - Component analysis: difference cases using different features , different feature extraction models and feature combinations for sentiment query answering.
  - Analysis:
    - **Only using VLM to extract visual feature to match sentiments performs best.**
    - **Data limitation makes simple classifier training can not be that effective.**

	All	Bagman	Manos	RoadToBali	TheIllusionist
V_{I3D}+T_{Bert}	0.19	0	0.25	0.25	0.25
V_{VLM}+T_{Bert}	0.19	0	0.25	0.25	0.25
V_{VLM}+T_{VLM}	0.19	0	0.25	0.25	0.25
V_{VLM}	**0.44**	**0.25**	**0.5**	**0.5**	**0.5**
T_{VLM}	0.25	**0.25**	0.25	0.25	0.25

V: visual feature; T: text feature; I3D: visual feature extraction model I3D; Bert: text feature extraction model BERT; VLM: visual and text feature extraction model VLM

Leaderboard

- **Movie-Level**
  - Group1: no rank
  - Group 2: rank 1
- **Scene-Level**
  - Group1: rank 3
  - Group 2: rank 3

Movie-Level Team Rank

Group	Rank 1	Rank 2	Rank 3
Group 1	UZH	N/A	N/A
Group 2	Nanjing Univ.	HERO TVQA	UZH

Scene-Level Team Rank

Group	Rank 1	Rank 2	Rank 3	Rank 4	Rank 5
Group 1	E-VG	Graphen	Nanjing Univ.	HERO TVQA	UZH
Group 2	DVU-SQL	HERO TVQA	Nanjing Univ.	Graphen	N/A

THANK YOU

MAGUS
MediA recoGnition
and UnderStanding