Deep Video Understanding with Video-Language Model

Runze Liu[1], Yaqun Fang[1], Fan Yu[1], Ruiqi Tian[2], Tongwei Ren[1,*], Gangshan Wu[1]

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

[2]University of British Columbia

Introduction

- **Deep Video Understanding Challenge (2020 ~ 2023)**
  - requires systems to develop a deep analysis and understanding of long video
  - use known information to reason about more hidden information
- **HLVU dataset**
  - 25 videos
    - 19 for development
    - 6 for test
  - Video length
    - 18 min – 110 min
    - 77 min in average
  - Annotations
    - scene, entity name, entity type, screenshot

Training dataset:

Honey – 86 min
Let's Bring Back Sophie – 50 min
Nuclear Family – 28 min
Shooters – 41 min
Spiritual Contact The Movie – 66 min
Super Hero – 18 min
The Adventures of Huckleberry Finn – 106 min
The Big Something – 101 min
Time Expired – 92 min
Valkaama – 93 min
Bagman – 107 min
Manos – 73 min
Road to Bali – 90 min
The Illusionist – 109 min
Chained for Life – 88 min
Liberty Kid – 88 min
Calloused Hands – 92 min
Like Me – 79 min
Losing Ground – 81 min

Testing dataset:

Achipelago – 110 min
Bonneville – 92 min
Heart Machine – 83 min
Little Rock – 82 min
Memphis – 78 min

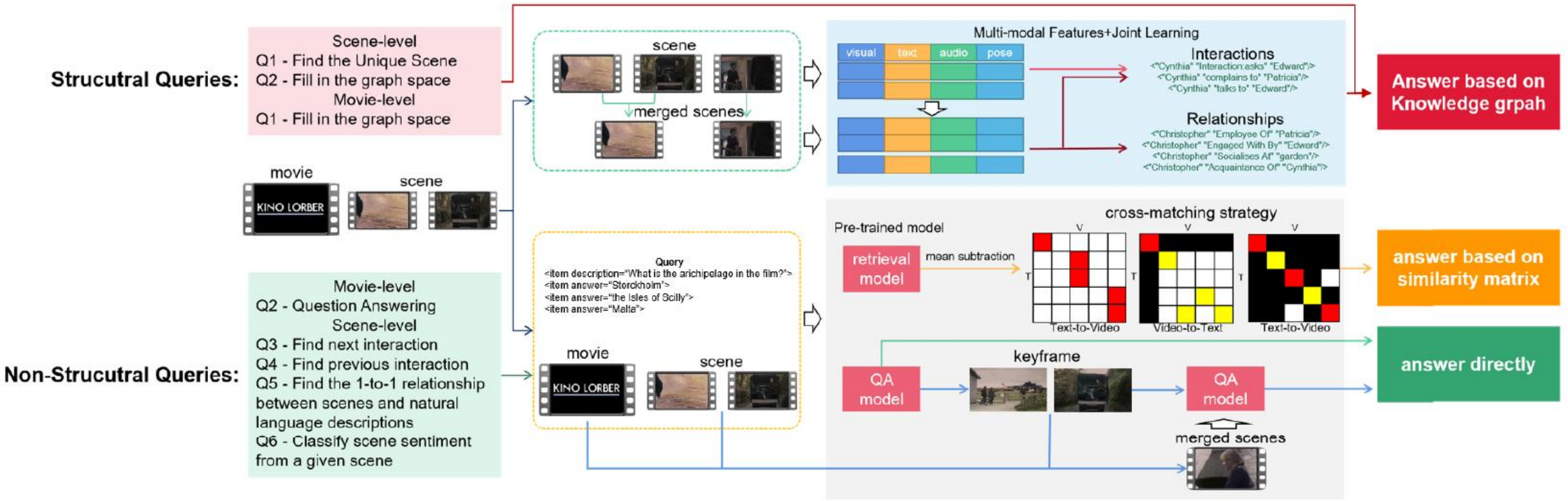
Task

- **Movie-level query types**
  - **Group 1**
    - Fill in the graph space
  - **Group 2**
    - Question Answering
- **Scene-level query types**
  - **Group 1**
    - Find the Unique Scene
    - Fill in the graph space
    - Find the next interaction
    - Find the previous interaction
  - **Group 2**
    - Find the 1-to-1 relationship between scenes and natural language descriptions
    - Classify scene sentiments from a given scene

Difficulties

- **Task**
  - All the tasks are based on high level semantic
  - There are significant differences between different tasks
  - Questions described by natural language added this year exceed the capabilities of knowledge graphs
- **Video**
  - Videos have multi-modal features
  - Movies last too long time
- **Text**
  - The semantic gap between some texts (such as predicate of queries) is very vague

Framework of our work



Structural Queries

- Multi-modal Feature Extraction
  - Visual feature, Text feature, Audio feature, Pose feature
- Merge scenes with LGSS
- Knowledge Graph Construction
  - Original scenes used for interaction knowledge graphs
  - Merged scenes used for relationship knowledge graphs
- Knowledge Graph Traversing
  - Scene-level: Traverse interaction knowledge graphs
  - Movie-level: Traverse relationship knowledge graphs

Scene-level
Q1 – Find the Unique Scene
Q2 – Fill in the graph space
Movie – level
Q1 – Fill in the graph space

Non-Structural Queries

- Retrieval Model
  - Scene-level Q5 Q6
- Question Answering (QA) Model
  - Movie-level Q2
  - Scene-level Q3 Q4
- Mean subtraction and cross-matching strategy
  - Scene-level Q5
- Input selection strategy
  - Movie-level Q2

Movie-level
Q2 – Question Answering
Scene-level
Q3 – Find next interaction
Q4 – Find previous interaction
Q5 – Find the 1-to-1 relationship between scenes and natural language descriptions
Q6 – Classify scene sentiment

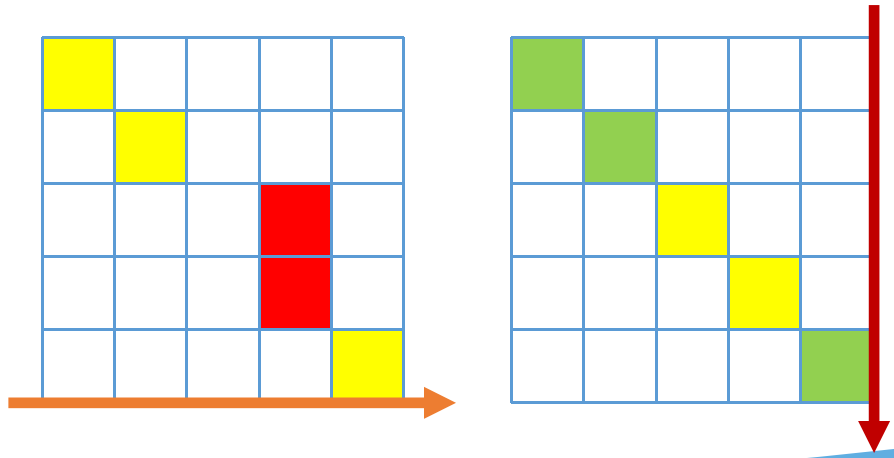
Non-Structural Queries Answer Inferring

- Mean subtraction and cross-matching strategy
  - Scene-level Q5
  - Q5 is a 10-to-10 query, not 10 * 1-to-10 query
  - Description lengths are similar, but scene lengths vary widely (about 10s ~ over 10 min)
  - Repeat match accuracy is up to 50%

Movie-level
Q2 – Question Answering
Scene-level
Q3 – Find next interaction
Q4 – Find previous interaction
Q5 – Find the 1-to-1 relationship between scenes and natural language descriptions
Q6 – Classify scene sentiment

	Scene A (10 s)	Scene B (10 min)
Talk	1 (match)	1 (cover)
Talk and Walk	0 (not match)	1 (match)

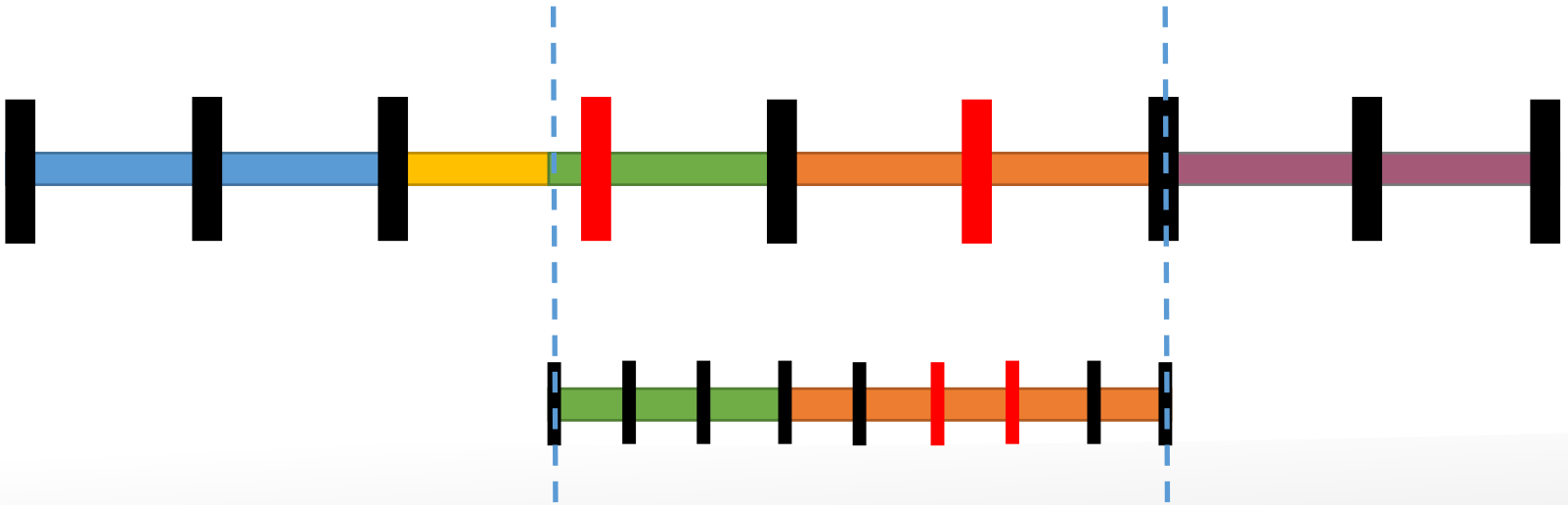
	Scene A (10 s)	Scene B (10 min)
Talk	0	0
Talk and Walk	-0.5	0.5



Non-Structural Queries Answer Inferring

- Input selection strategy
  - Movie-level Q2
  - Compress movie from about 25FPS to 3FPS
  - QA model used (SeViLA) provides keyframes as a basis for answering question
  - Select scenes according to keyframes and use QA model to answer the same query again

Movie-level
Q2 – Question Answering
Scene-level
Q3 – Find next interaction
Q4 – Find previous interaction
Q5 – Find the 1-to-1 relationship between scenes and natural language descriptions
Q6 – Classify scene sentiment



Experiment

- **Description-and-scene retrieval**
  - Metric: accuracy
  - Analysis:
    - Clip4clip performs better than XClip
    - Using the combination of mean subtraction and cross-strategy performs best

	XClip					*Clip4clip*				
	$V2T$	$V2T_{MS}$	$T2V$	$T2V_{CS}$	$T2V_{MS+CS}$	$V2T$	$V2T_{MS}$	$T2V$	$T2V_{CS}$	$T2V_{MS+CS}$
CallousedHands	0.50	0.40	0.50	0.60	0.60	0.70	0.60	0.60	0.60	**0.70**
ChainedforLife	0.60	0.70	0.50	**0.70**	**0.70**	0.70	0.60	**0.70**	**0.70**	**0.70**
LibertyKid	0.70	0.80	0.70	0.80	**1.00**	0.90	0.90	0.90	0.90	0.90
LikeMe	0.60	0.60	0.60	0.70	0.60	0.80	0.80	0.70	0.70	**0.80**
LosingGround	0.20	0.50	0.40	0.40	0.50	0.40	0.60	0.50	**0.70**	**0.70**
All	0.52	0.60	0.54	0.64	0.68	0.70	0.70	0.68	0.72	**0.76**

Experiment

- **Next-and-Previous Interaction Prediction**
  - Metric: accuracy
  - Analysis:
    - More keyframes (KF) performs better
    - Question Answering (QA) model performs better than Retrieval model

	Next-and-Previous Interaction Prediction (Q3 and Q4)				
	$XClip_{V2T}$	$Clip4clip_{V2T}$	$QA_{KF=1}$	$QA_{KF=2}$	$QA_{KF=4}$
CallousedHands	**0.25**	**0.25**	**0.25**	**0.25**	**0.25**
ChainedforLife	0.25	0.25	0.25	0.25	**0.38**
LibertyKid	0.13	0.00	0.25	0.25	**0.38**
LikeMe	0.13	**0.38**	0.25	0.25	0.25
LosingGround	0.13	0.13	**0.75**	**0.75**	**0.75**
All	0.18	0.20	0.35	0.35	**0.40**

Experiment

- **Sentiment Retrieval**
  - Metric: accuracy
  - Analysis:
    - Less keyframes (KF) performs better
    - Retrieval model performs better than Question Answering (QA) model

	Sentiment Retrieval (Q6)				
	$XClip_{V2T}$	$Clip4clip_{V2T}$	$QA_{KF=1}$	$QA_{KF=2}$	$QA_{KF=4}$
CallousedHands	0.50	0.50	**0.67**	**0.67**	**0.67**
ChainedforLife	**0.67**	**0.67**	0.33	0.33	0.17
LibertyKid	**0.50**	**0.50**	**0.50**	0.33	0.33
LikeMe	**0.50**	**0.50**	0.17	0.17	0.17
LosingGround	**0.50**	0.17	**0.50**	**0.50**	0.33
All	**0.53**	0.47	0.43	0.40	0.33

Leaderboard

- **Movie-level**
  - Group 1: Rank 4
  - Group 2: Rank 2

Level	Group	Rank 1	Rank 2	Rank 3	Rank 4
Movie	Group1	STARS	WHU_ NERCMS	MINE-MM	**MAGUS.LFYT**
	Group2	WHU_ NERCMS	**MAGUS.LFYT**	MINE-MM	N/A
Scene	Group1	WHU_ NERCMS	**MAGUS.LFYT**	MINE-MM	N/A
	Group2	**MAGUS.LFYT**	WHU_ NERCMS	DSSC	MINE-MM

- **Scene-level**
  - Group 1: Rank 2

THANK YOU

MAGUS

MediA recoGnition
and UnderStanding