JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

A-CAP: Anticipation Captioning with Commonsense Knowledge

Duc Minh Vo[1*] Quoc-An Luong[2] Akihiro Sugimoto[2] Hideki Nakayama[1]

[1]The University of Tokyo, Japan [2]National Institute of Informatics, Japan

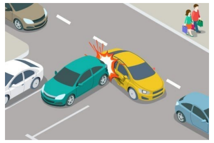
*Presenter Paper

WED-AM-247

HIGHLIGHTS

- Why is the output caption?
  - Inherit from success of image captioning
  - Flexible transformation

- Applications:



Incident prevention

Falling prediction

Crime prevention

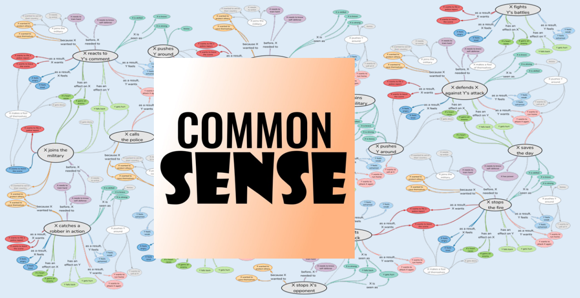
NEW

ANTICIPATION CAPTIONING

Generate caption for unseen image that is the future of seen images

HIGHLIGHTS

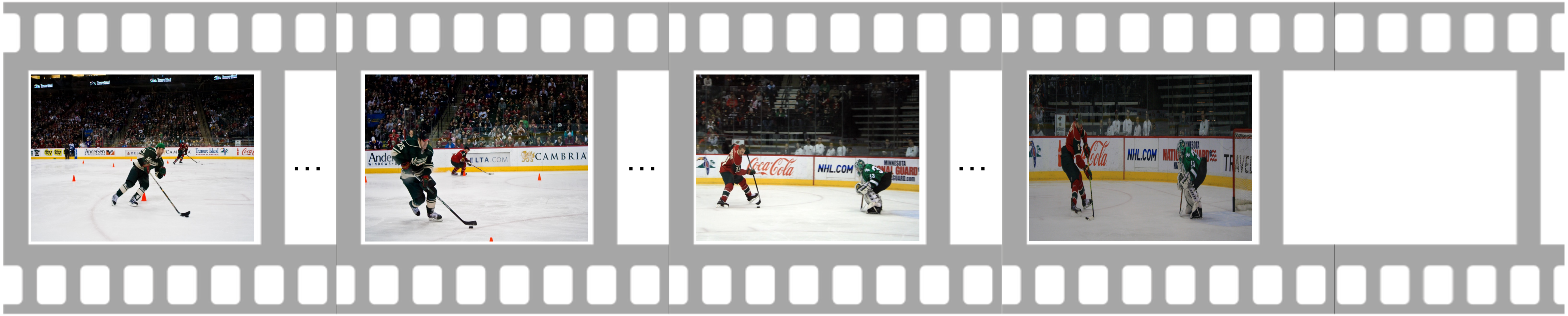
A-CAP



He shoots, he scores and the game ends one to nothing.

DEFINITION OF ANTICIPATION CAPTIONING TASK

(seen) Sparsely temporally-ordered images **(unseen) Oracle image**



Input **Output**

- “sparse’” means that **two consecutive images** are not required to be as close in time as those in a video
- the potential future may be oracle image (for reference only)

Caption for the unseen oracle image

OUTPUTS OF DIFFERENT TASKS

(seen) Sparsely temporally-ordered images **(unseen) Oracle image**

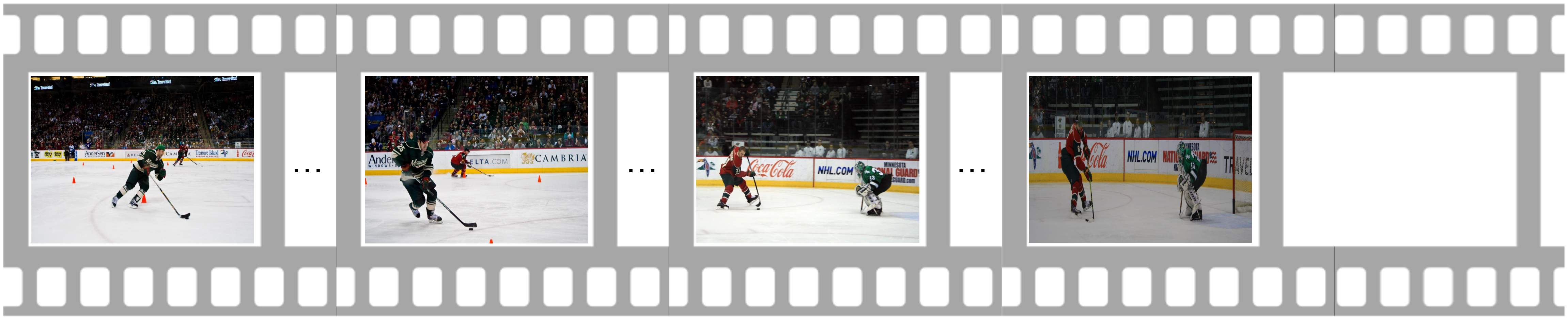
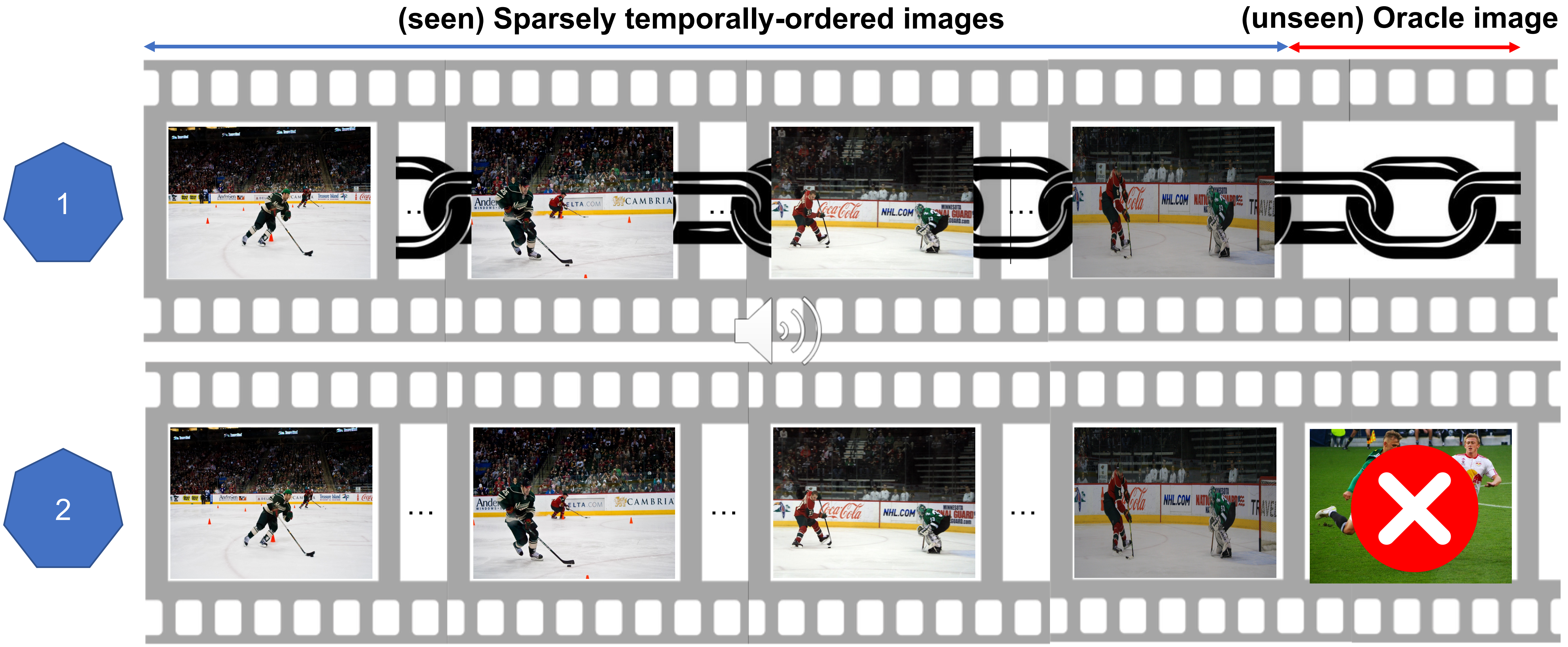


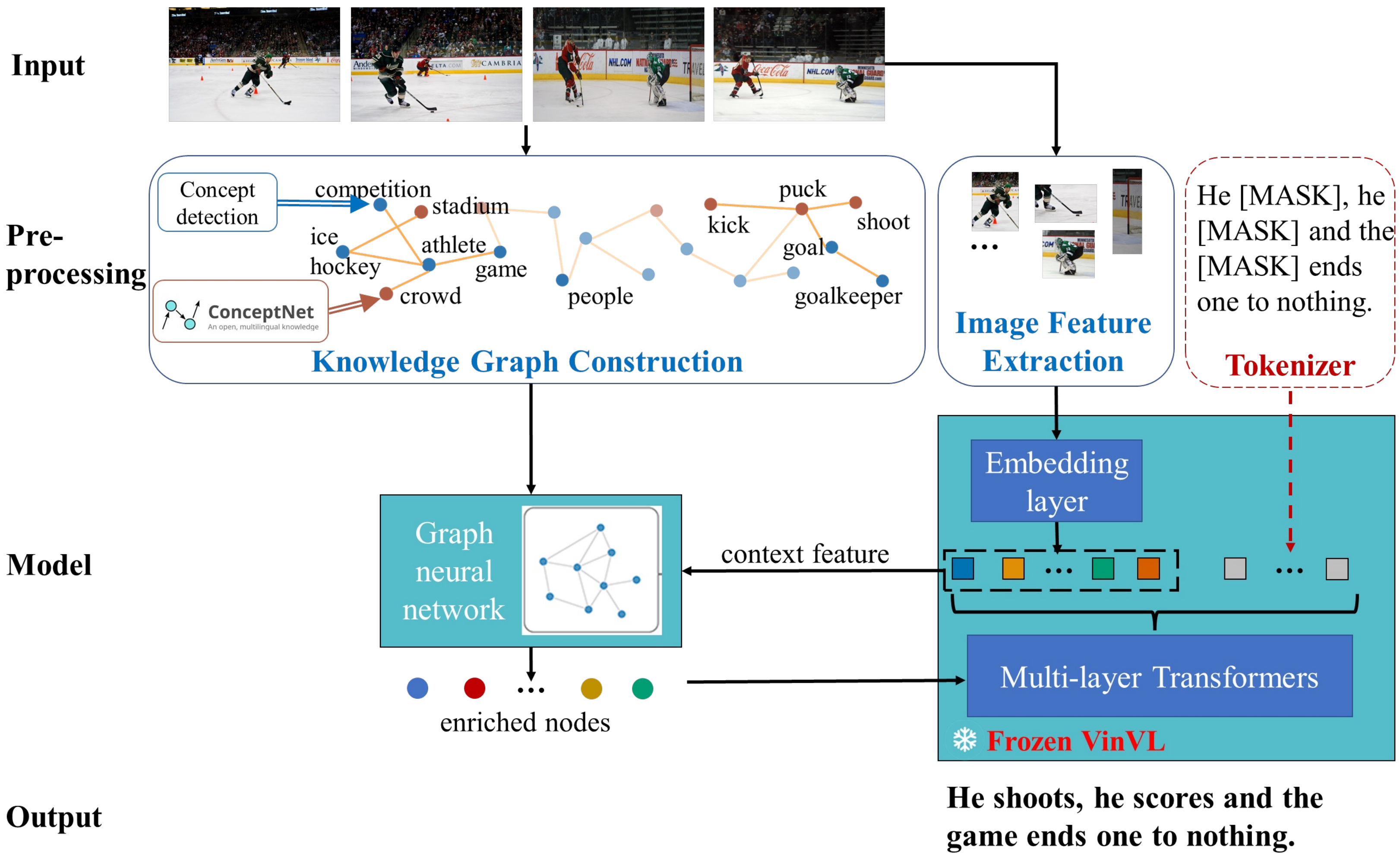
Image captioning	A man standing with a hockey stick.	A man standing on an ice rink.	A man holding a hockey stick.	A group of men playing a game of hockey.	
Story telling	This breakaway was the first threat to score. The wingman took the puck to the goal but a nice play by the goalie saved the goal. Finally, the other team gets the puck deep into red zone. He is now within 20 feet of the goal.				
Anticipation captioning					He shoots, he scores and the game ends one to nothing.

OUR HYPOTHESES



Use commonsense knowledge to connect all detected concepts while retrieving forecasted ones, creating a knowledge graph

OUR PROPOSED A-CAP MODEL



- Construct *knowledge graph* using concept detection and ConcepNet
- Extract image features
- Graph network to enrich nodes, average of image features is used context feature
- *Frozen* vision - language VinVL
- Cross-entropy loss

DATASET AND COMPARED METHODS

- We customize the Visual storytelling dataset (VIST)

Original VIST

Input


Ground-truth output

This breakaway was the first threat to score. The wingman took the puck to the goal but a nice play by the goalie saved the goal. Finally, the other team gets the puck deep into red zone. He is now within 20 feet of the goal. **He shoots, he scores and the game ends one to nothing.**

Our dataset

Input


Ground-truth output

Oracle image

Ting-Hao K. Huang et al. Visual storytelling. In NAACL, 2016

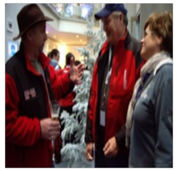
DATASET AND COMPARED METHODS

- We customize the Visual storytelling dataset (VIST)

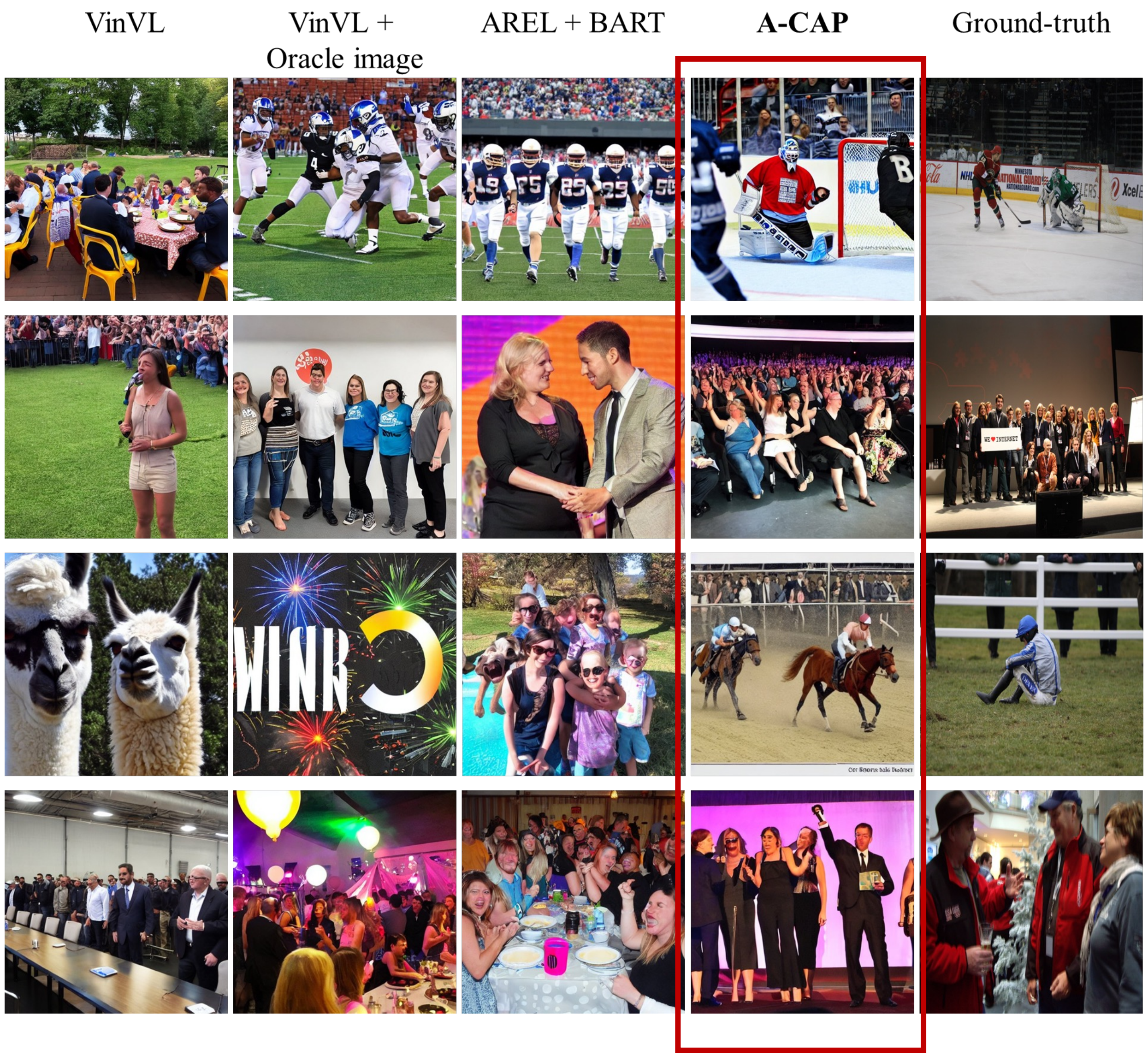
- Compared methods:
  - VinVL: image captioning model. We replace its single image input by the sequence of images
  - VinVL+Oracle image is the method where VinVL uses the ground-truth oracle image in training and testing
  - AREL+ BART is a combination of visual storytelling and story ending generation

Pengchuan Zhang et al. Vinvl: Making visual representations matter in vision-language models. In CVPR, 2021.
Xin Wang et al. No metrics are perfect: Adversarial reward learning for visual storytelling. In ACL, 2018.
Mike Lewis et al. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In ACL, 2020.

QUALITATIVE COMPARISON

Sparsely temporally-ordered images	Oracle image (for reference)	VinVL	VinVL + Oracle image	AREL + BART	**A-CAP**	Ground-truth
		after the ceremony, the teams got to eat outside. Out of context	the defense was able to close out the game and had a great time. Sometimes reasonable	i was getting ready to leave the game and i took a picture of the players on the field before the game.	the goalie caught the puck as it passed the goalie. More plausible	he shoots, he scores and the game ends one to nothing.
		she let the crowd ask questions in the end.	we got to meet the people behind the company's logo.	he welcomed to the stage his new assistant	at the end of the show, the audience enjoyed themselves.	they were all about preserving the internet
		the llamas were very curious.	the competition ended with a bang.	they had a great time. General ending	it was a great time for the horse racers.	he thought he was going to cry
		the vice president closed the meeting by thanking all the workers of the company.	the party went on well into the night.	everyone was having a great time.	they ended the night with a speech.	eventually the winner was announced, and he was very grateful

QUALITATIVE COMPARISON



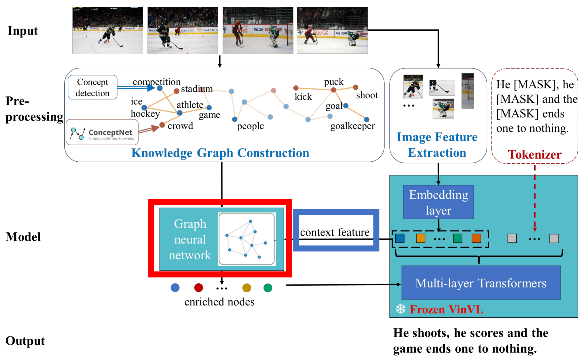
- We use stable diffusion to generate images using the anticipated captions in previous slide.
- Our generated images are closed to the ground-truth ones while those by other methods are not.

QUANTITATIVE COMPARISON

Method	Accuracy ↑						Descriptiveness ↑		
	B-1	B-4	CIDEr	SPICE	CLIPScore	RefCLIPScore	R@1	R@5	R@10
VinVL	31.7	3.1	2.6	13.8	40.7	42.8	1.3	6.5	10.8
VinVL + Oracle image	34.9	3.8	4.3	16.9	57.9	61.3	8.1	17.2	31.1
AREL + BART	30.9	2.0	3.1	11.4	37.8	39.7	1.1	5.9	9.3
A-CAP	**37.2**	**6.9**	**4.7**	**20.1**	**65.2**	**70.2**	**8.7**	**18.9**	**31.5**
Δ	2.3↑	3.1↑	0.4↑	3.2↑	7.3↑	8.9↑	0.6↑	1.7↑	0.4↑

Our method outperforms others on all metric

ABLATION STUDY



Method	Accuracy						Descriptiveness		
	B-1	B-4	CIDEr	SPICE	CLIPScore	RefCLIPScore	R@1	R@5	R@10
A-CAP	**37.2**	**6.9**	**4.7**	**20.1**	**65.2**	**70.2**	**8.7**	**18.9**	**31.5**
A-CAP w/o GNN	34.8	5.2	3.7	14.5	38.2	47.3	3.6	8.7	15.4
A-CAP w/o context	36.1	6.2	4.2	13.9	39.8	46.9	4.1	9.5	16.1

The performance scores by ablated models are degraded

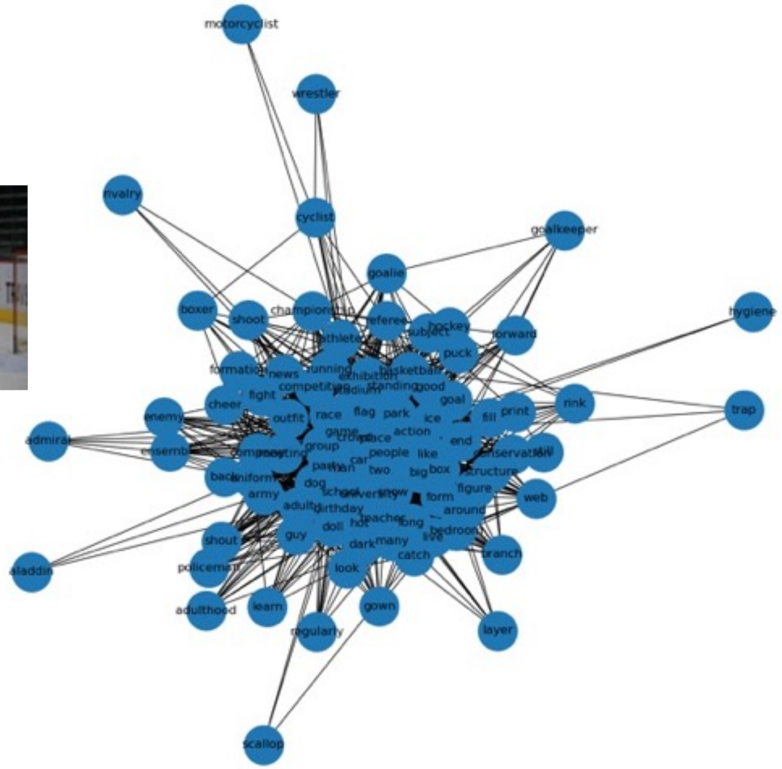
Sparsely temporally-ordered images	Oracle image (for reference)	A-CAP w/o GNN	A-CAP w/o context	**A-CAP** (full)	Ground-truth
		the downtown streets were lined with people enjoying themselves	we walked around a bit more before heading home.	we ended the night by shopping in the center of the city.	we stopped at a souvenir store to get some things before finally heading back to the hotel.
		the nightlife was just amazing to look at.	we had a great time walking around.	it was a great night and i can't wait to go back next year.	the place was ready to close and we had to leave.

- We select two inputs where the detected concepts almost overlap.
- A-CAP w/o GNN generates captions that most likely describe the inputs.
- A-CAP w/o context generates captions that are far from the inputs and similar to each other.

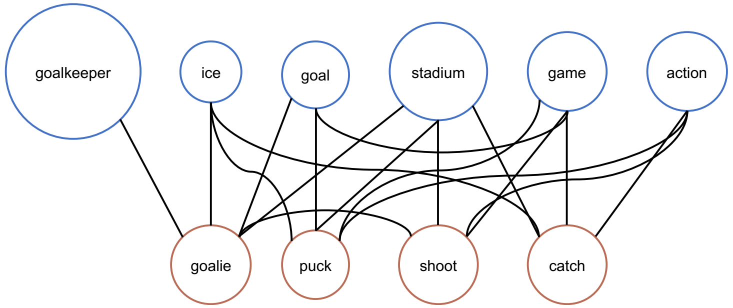
EXAMPLE OF KNOWLEDGE GRAPH

Input

Output The goalie caught the puck as it passed the goalie.



Full knowledge graph



The detected concepts and forecasted concepts that contribute to the output

FAILED CASE

Sparsely temporally-ordered images	Oracle image (for reference)	**A-CAP**	Ground-truth
		the bride and groom are about to cut the cake.	night settles on this wonderous day and everyone heads home.

The reason is that the oracle image changes significantly from the inputs

THANK YOU FOR YOUR ATTENTION

SEE YOU AT POSTER SESSION WED-AM-247