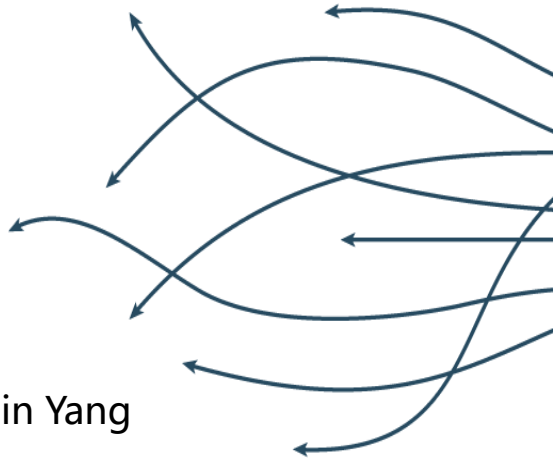
Vision-Language Learning and Beyond

Yadong Mu
Wangxuan Institute of Computer Technology
Peking University

Joint work with my graduate students
Gong Guoqiang, Bao Peijun, Yongzhi Li, Zhou Xinzhe, Jin Yang



This talk is a wrap-up of the following papers

- [1] Lu Chi, Guiyu Tian, Yadong Mu, Qi Tian, Two-Stream Video Classification with Cross-Modality Attention, ICCV Workshops 2019.
- [2] Guiyu Tian, Shuai Wang, Jie Feng, Li Zhou, Yadong Mu, Cap2Seg: Inferring Semantic and Spatial Context from Captions for Zero-Shot Image Segmentation, ACM International Conference on Multimedia (ACMMM) 2020.
- [3] Guoqiang Gong, Liangfeng Zheng, Kun Bai, Yadong Mu, Scale Matters: Temporal Scale Aggregation Network for Precise Action Localization in Untrimmed Videos, IEEE International Conference on Multimedia and Expo (ICME) 2020 (Oral Presentation)
- [4] Xinzhe Zhou, Yadong Mu, Google Helps YouTube: Learning Few-Shot Video Classification from Historic Tasks and Cross-Domain Sample Transfer, ACM International Conference on Multimedia Retrieval (ICMR) 2020 (Oral Presentation)
- [5] Ruihai Wu, Kehan Xu, Chenchen Liu, Nan Zhuang, Yadong Mu, Localize, Assemble, and Predicate: Contextual Object Proposal Embedding for Visual Relation Detection, Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI-2020) (Oral Presentation)
- [6] Baifeng Shi, Qi Dai, Yadong Mu, Jingdong Wang, Weakly-Supervised Action Localization by Generative Attention Modeling, IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2020.
- [7] Guoqiang Gong, Xinghan Wang, Yadong Mu, Qi Tian, Learning Temporal Co-Attention Models for Unsupervised Video Action Localization, IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2020 (Oral Presentation).
- [8] Yongzhi Li, Duo Zhang, Yadong Mu, Visual-Semantic Matching by Exploring High-Order Attention and Distraction, IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2020.
- [9] Chenchen Liu, Yang Jin, Kehan Xu, Guoqiang Gong, Yadong Mu, Beyond Short-Term Snippet: Video Relation Detection with Spatio-Temporal Global Context, IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2020.

PDF-version manuscripts can be accessed at http://www.muyadong.com/publication.html

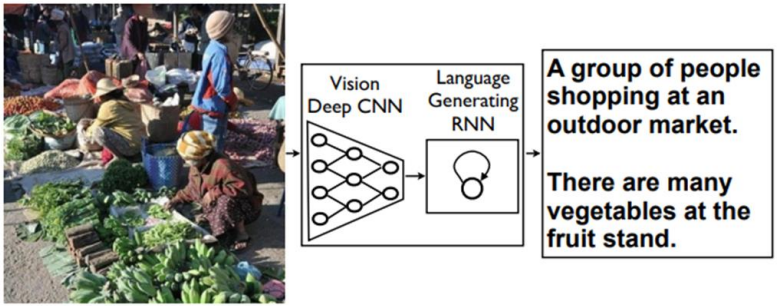
This talk is a wrap-up of the following papers

- [10] Yongzhi Li, Yadong Mu, Nan Zhuang, Xianglong Liu, Efficient Fine-Grained Visual-Text Search Using Adversarially-Learned Hash Codes, IEEE International Conference on Multimedia and Expo (ICME) 2021.
- [11] Xinzhe Zhou, Yadong Mu, Question-Guided Semantic Dual-Graph Visual Reasoning with Novel Answers, The Annual ACM International Conference on Multimedia Retrieval (ICMR) 2021.
- [12] Guoqiang Gong, Liangfeng Zheng, Wenhao Jiang, Yadong Mu, Self-Supervised Video Action Localization with Adversarial Temporal Transforms, The 30th International Joint Conference on Artificial Intelligence (IJCAI) 2021.
- [13] Yang Jin, Wenhao Jiang, Yi Yang, Yadong Mu, Zero-Shot Video Event Detection with High-Order Semantic Concept Discovery and Matching, IEEE Transactions on Multimedia, 2021.
- [14] Peijun Bao, Qian Zheng, Yadong Mu, Dense Events Grounding in Video, Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI) 2021.
- [15] Xinzhe Zhou, Wei Liu, Yadong Mu, Rethinking the Spatial Route Prior in Vision-and-Language Navigation, under review
- [16] Peijun Bao, Yadong Mu, Learning Sample Importance for Cross-Scenario Video Temporal Grounding, under review

PDF-version manuscripts can be accessed at
http://www.muyadong.com/publication.html

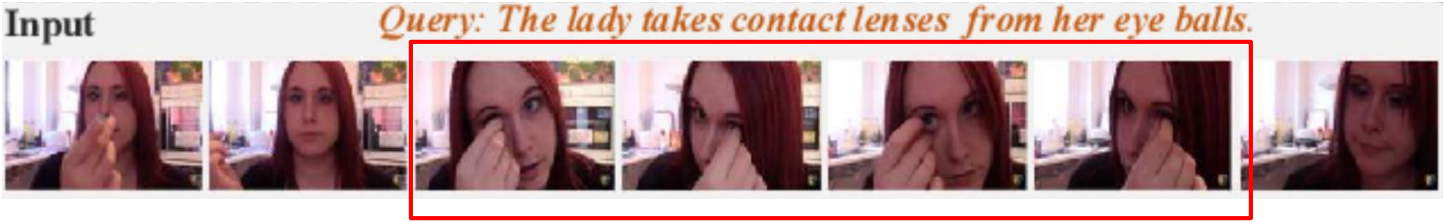
Vision and Language

Captioning



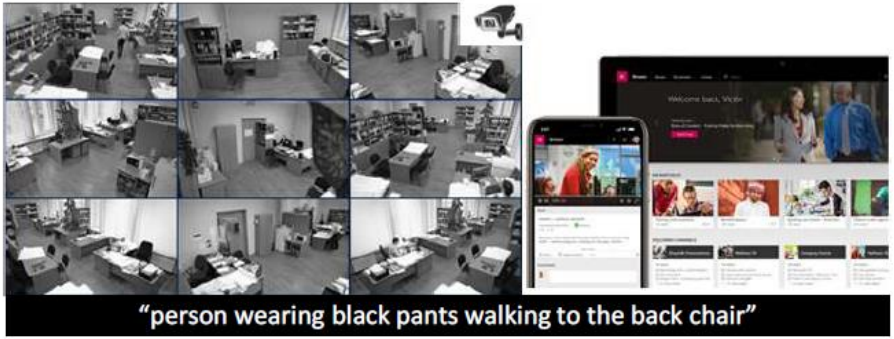
Grounding

(Spatio)-Temporal Localization

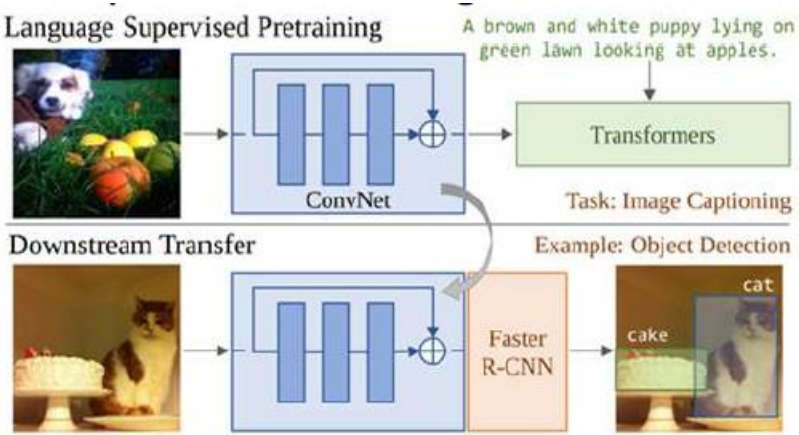


Why Language in Vision?

Human-computer interaction



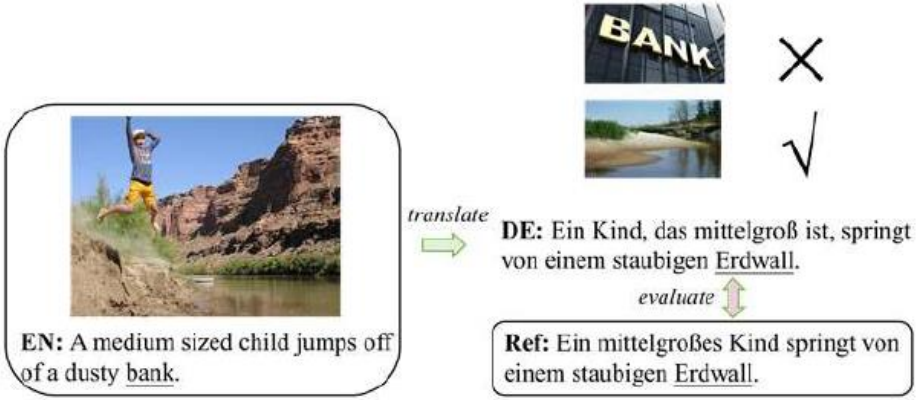
Language-based search



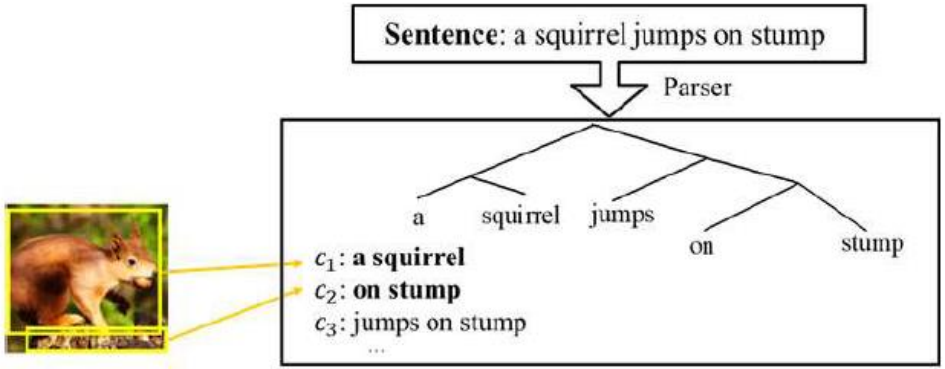
Visual representation learning with language supervision

Slides credit to Prof. Jiebo Luo

Why Vision in Language?



Multimodal machine translation



Unsupervised grammar induction

Slides credit to Prof. Jiebo Luo

A Collection of V-L Learning Tasks

Visual Captioning

A horse carrying a large load of hay and two people sitting on it.

train on the tracks. trees are green. front of the train is yellow. grass is green. green trees in the background. photo taken during the day. red train car.

- ***Popular Topics***: Advanced attentions, RL/GAN-based model training, Style diversity, Language richness, Evaluation
- ***Popular Tasks***: Image/video captioning, Dense captioning, Storytelling

Visual QA/Grounding/Reasoning

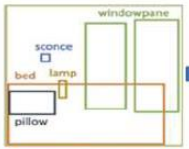
VQA

Referring Expressions

- ***Popular Topics***: Multimodal fusion, Advanced attentions, Use of relations, Neural modules, Language bias reduction
- ***Popular Tasks***: VQA, GQA, VisDial, Ref-COCO, CLEVR, VCR, NLVR2

Text-to-image Synthesis

This bird is red with white belly and has a very short beak



Popular Tasks:

- Text-to-image
- Layout-to-image
- Scene-graph-to-image
- Text-based image editing
- Story visualization

SOTA Models:

- StackGAN
- AttnGAN
- ObjGAN
- ...

Machine Translation/Grammar Induction

EN: A medium sized child jumps off of a dusty bank.

translate

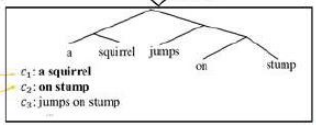
DE: Ein Kind, das mittelgroß ist, springt von einem staubigen Erdwall.

evaluate

Ref: Ein mittelgroßes Kind springt von einem staubigen Erdwall.

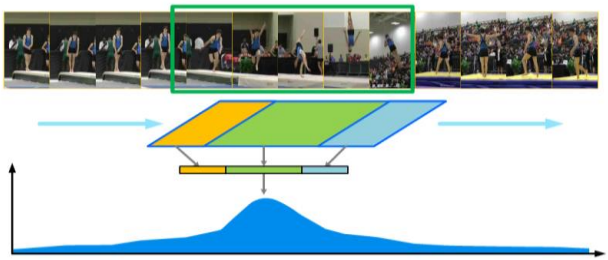
Sentence: a squirrel jumps on stump

Parser

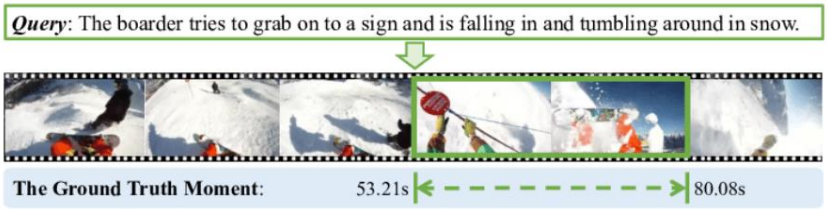


Slides credit to Prof. Jiebo Luo

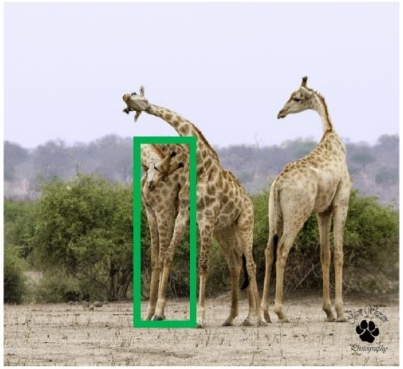
Visual Localization and Grounding



Temporal Action Localization



Moment Localization



RefCOCO:
1. giraffe on left
2. first giraffe on left

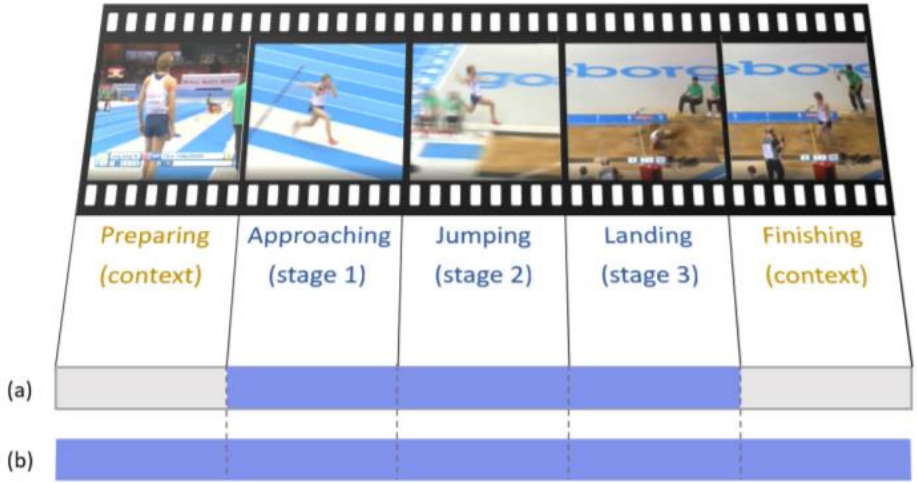
RefCOCO+:
1. giraffe with lowered head
2. giraffe head down

RefCOCOg:
1. an adult giraffe scratching its back with its horn
2. giraffe hugging another giraffe

Visual Grounding

Generative Attention for Weakly-Supervised TAL

- Weak-supervision: annotate an action with video-level labels
- The action-context confusion issue
- Our solution: take a discriminative-generative perspective



(a) ground-truth of action localization.
(b) Action-context confusion.

Baifeng Shi, Qi Dai, Yadong Mu, Jingdong Wang, Weakly-Supervised Action Localization by Generative Attention Modeling, CVPR 2020

Generative Attention for Weakly-Supervised TAL

- The original maximum-a-posteriori problem

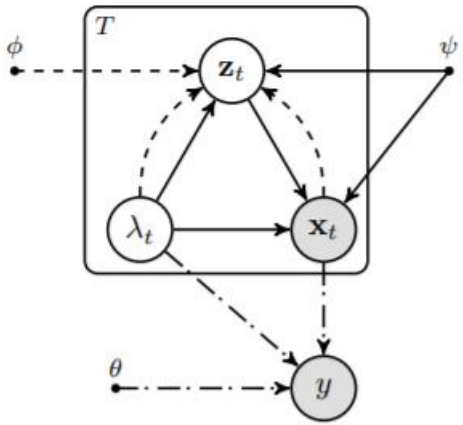
$$\max_{\lambda_t \in [0,1]} \log p(\boldsymbol{\lambda}|\mathbf{X}, y)$$

- Apply the Bayes' rule

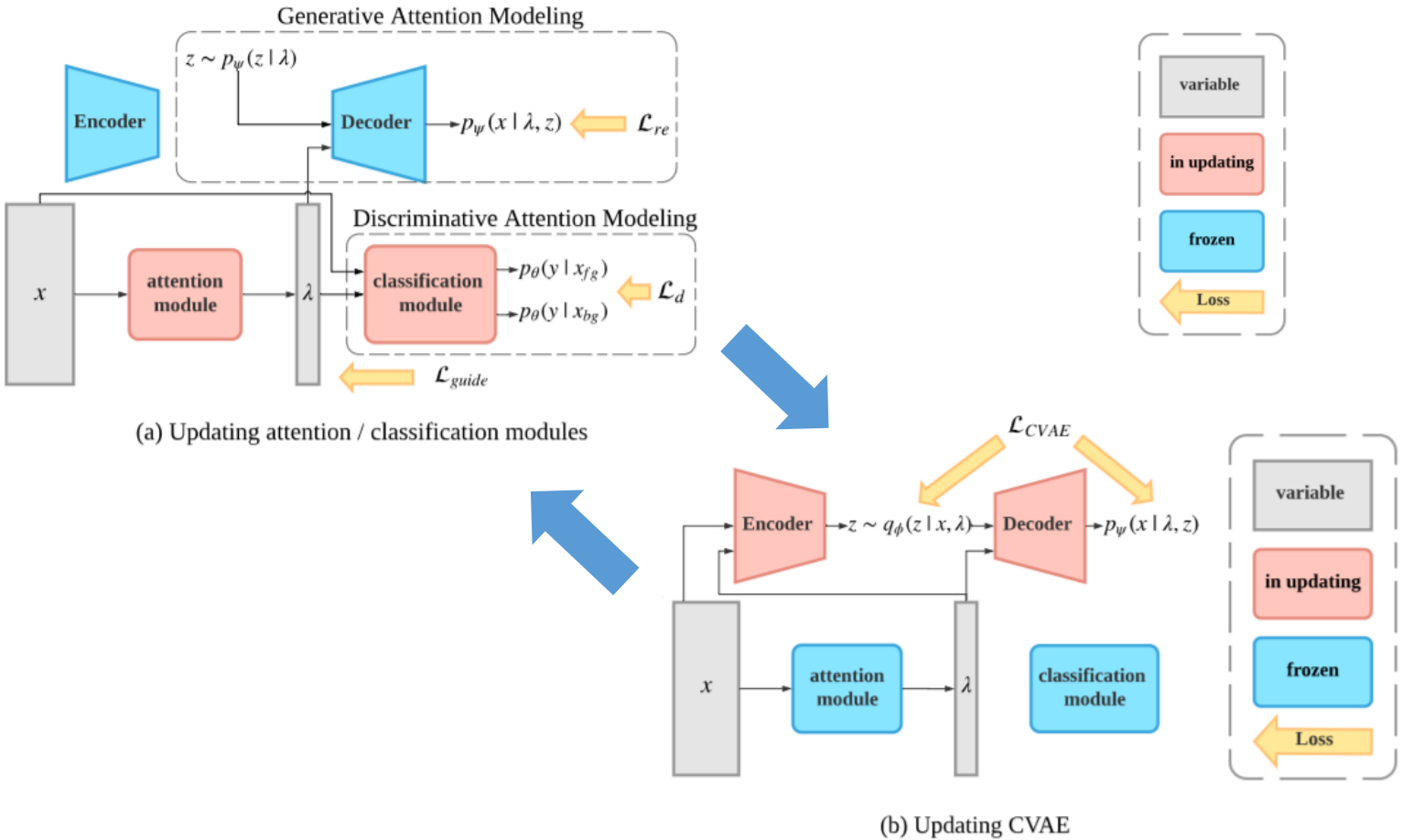
$$\begin{aligned}\log p(\boldsymbol{\lambda}|\mathbf{X}, y) &= \log p(\mathbf{X}, y|\boldsymbol{\lambda}) + \log p(\boldsymbol{\lambda}) - \log p(\mathbf{X}, y) \\ &= \log p(y|\mathbf{X}, \boldsymbol{\lambda}) + \log p(\mathbf{X}|\boldsymbol{\lambda}) + \log p(\boldsymbol{\lambda}) \\ &\quad - \log p(\mathbf{X}, y) \\ &\propto \log p(y|\mathbf{X}, \boldsymbol{\lambda}) + \log p(\mathbf{X}|\boldsymbol{\lambda}),\end{aligned}$$

discriminative

generative

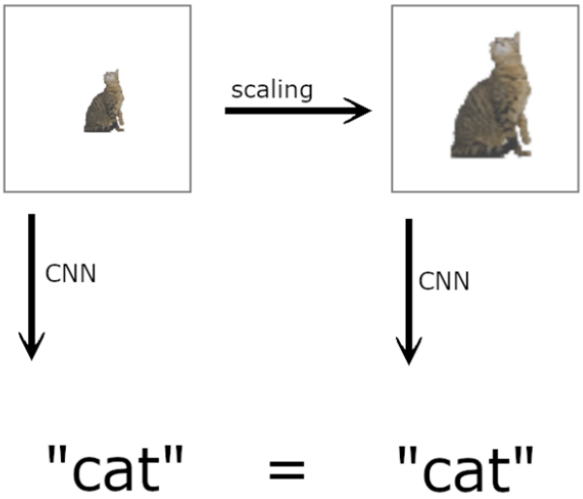


Generative Attention for Weakly-Supervised TAL

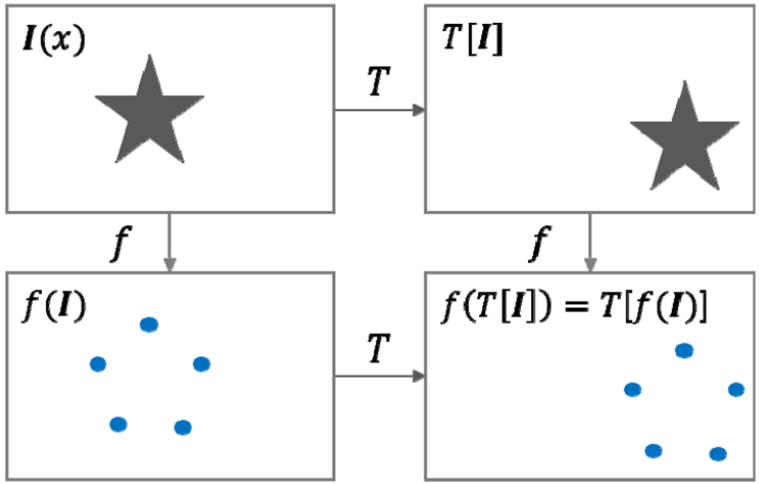


Equivariance-Induced Self-supervision in Weak-TAL

invariance

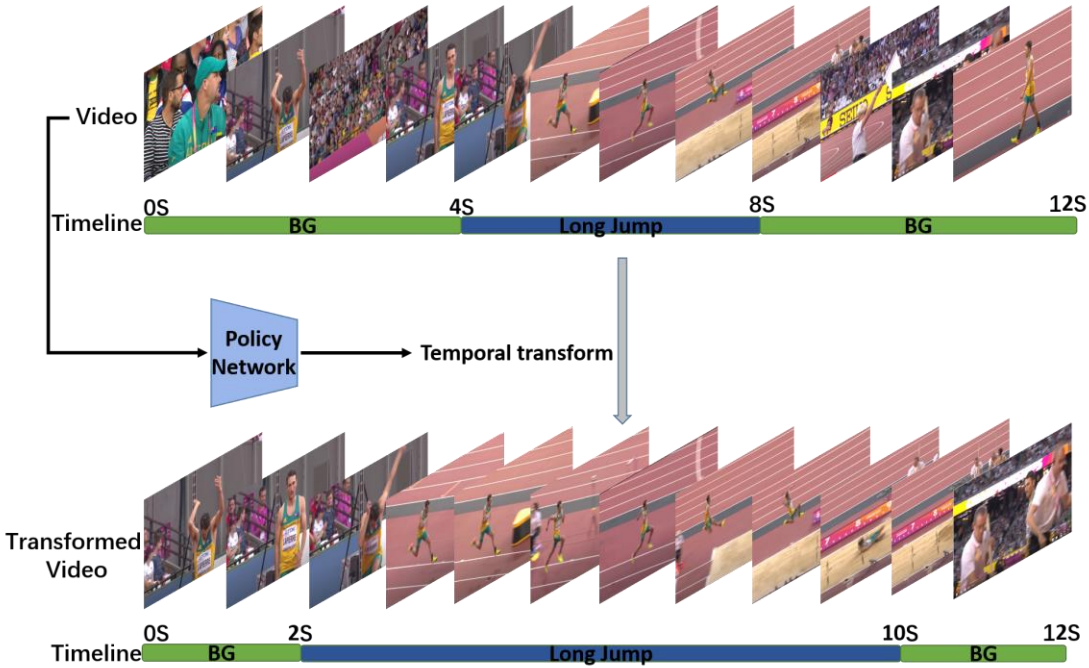


equivariance



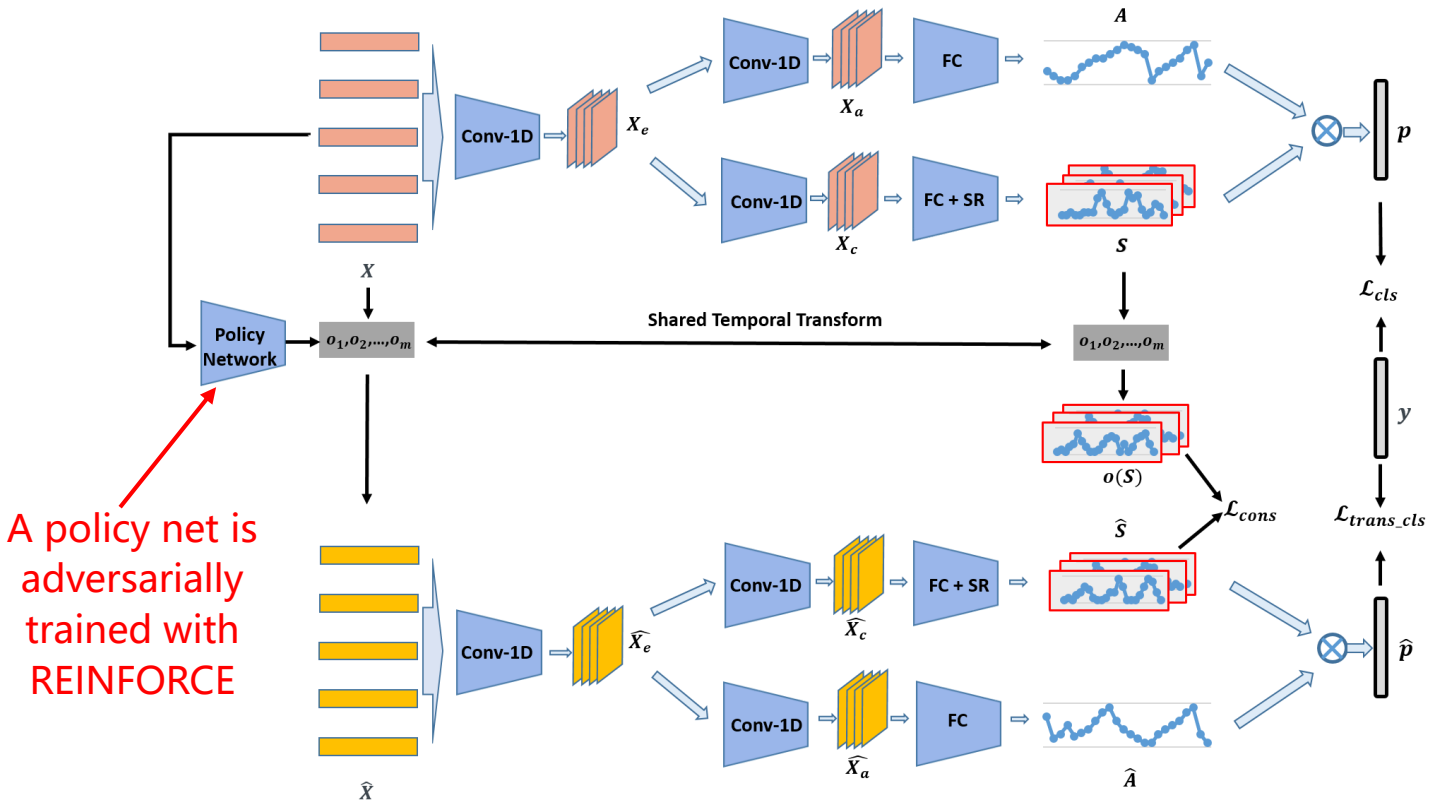
Equivariance-Induced Self-supervision in Weak-TAL

- When temporal transforms are applied
  - Video labels are invariant
  - Temporal boundaries are equivariant



Guoqiang Gong, Liangfeng Zheng, Wenhao Jiang, Yadong Mu, Self-Supervised Video Action Localization with Adversarial Temporal Transforms, The 30th International Joint Conference on Artificial Intelligence (IJCAI) 2021.

Equivariance-Induced Self-supervision in Weak-TAL

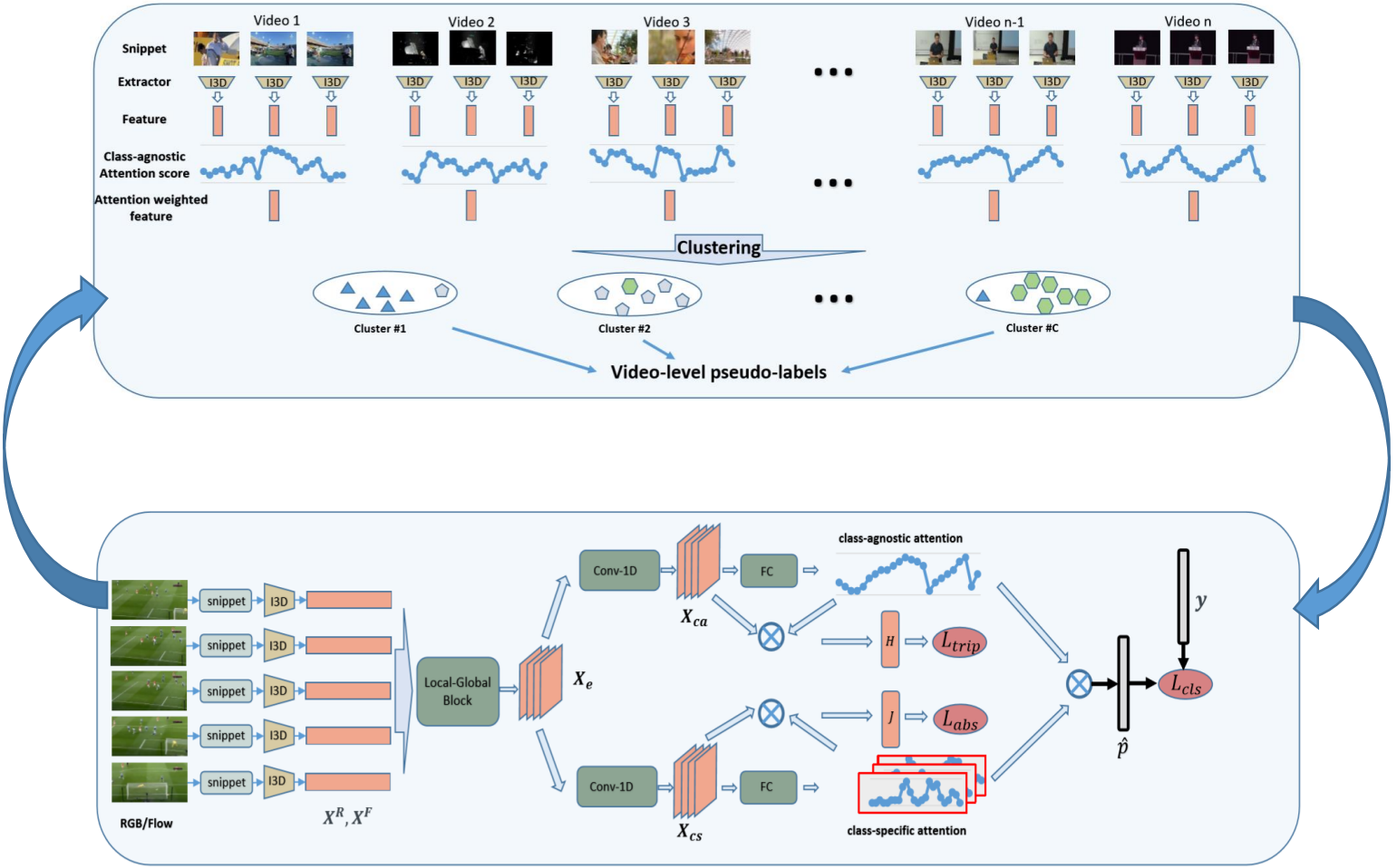


Unsupervised Temporal Action Localization

- Is TAL w/o any annotation technically possible?
- A Chicken (frame-level attention) and Egg (video-level label) problem

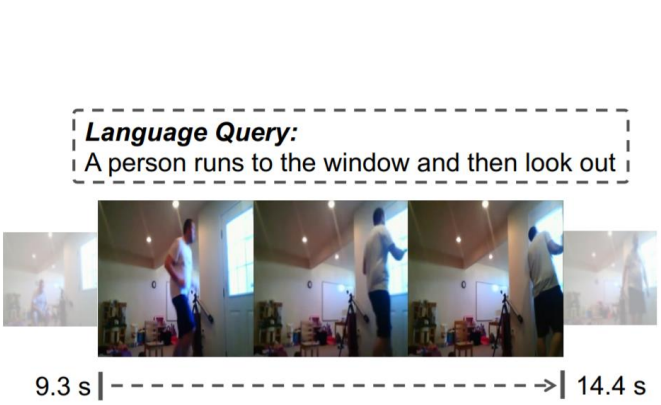
Guoqiang Gong, Xinghan Wang, Yadong Mu, Qi Tian, Learning Temporal Co-Attention Models for Unsupervised Video Action Localization, IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2020 (Oral Presentation).

The Proposed Alternating Procedure

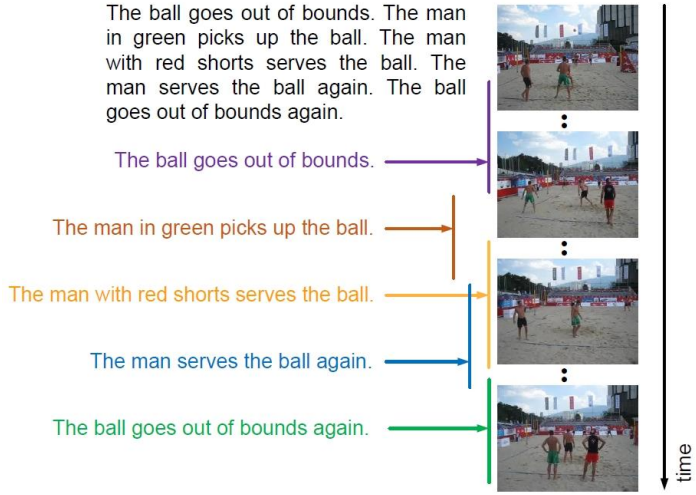


Dense Events Grounding

- Explore a novel setting of temporal grounding: dense events grounding (i.e. jointly localize multiple events described by a paragraph)



Grounding single event described by sentence

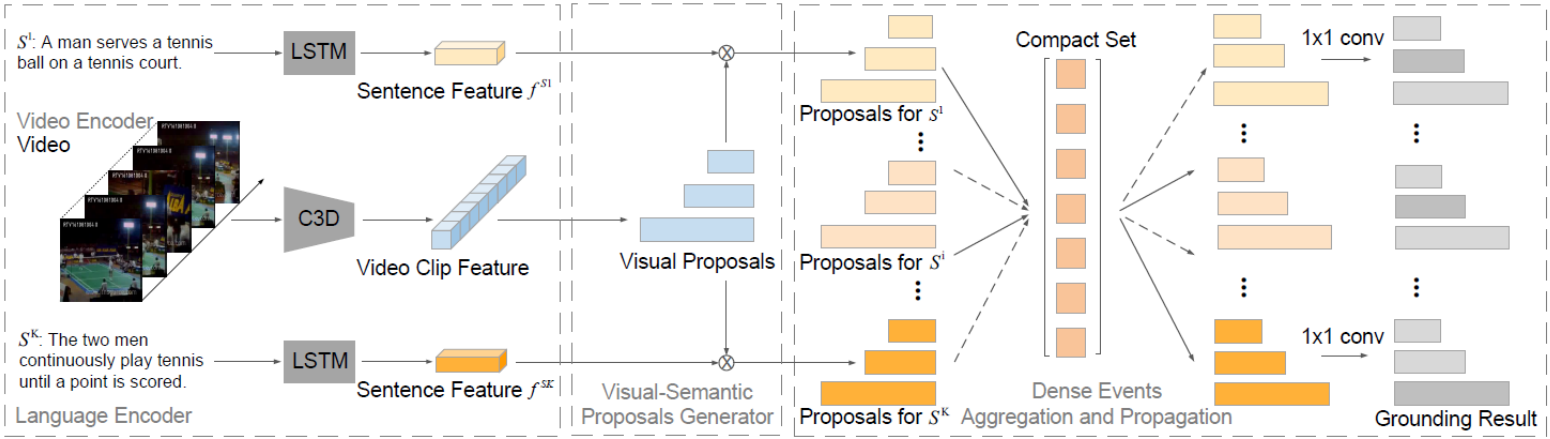


Grounding dense events described by paragraph

Peijun Bao, Qian Zheng, Yadong Mu, Dense Events Grounding in Video, AAAI 2021.

Dense Events Grounding

- Solution: Aggregation and Propagation Mechanism
  - Aggregate temporal and semantic information of dense events into a compact set
  - Selectively propagates the aggregated information to each single event



Dense Events Propagation Network (DepNet)

Cross-Scenario Temporal Grounding

A pilot experiment:

- zero all the sentence queries using only the video information

Surprisingly, we find its performance:

- significantly outstrips random guess, comparable to normal models

Table 1: Performance evaluation results on the ActivityNetCap.

Input	Method	R@1 IoU=0.5	R@1 IoU=0.7	R@5 IoU=0.5	R@5 IoU=0.7
	random	13.99	4.69	44.69	17.64
video & query	CTRL	29.01	10.34	59.17	37.54
	PFGA	33.04	19.26	-	-
	SCDM	36.75	19.86	64.99	41.53
	2D-TAN	44.51	26.54	77.13	61.96
	TLL	44.24	27.01	75.22	60.23
video-only	PFGA	21.69	12.56	-	-
	SCDM	23.84	12.93	51.66	32.36
	2D-TAN	27.56	13.93	61.65	36.78
	TLL	28.10	13.96	59.07	36.25

Models using only the visual information still perform well.

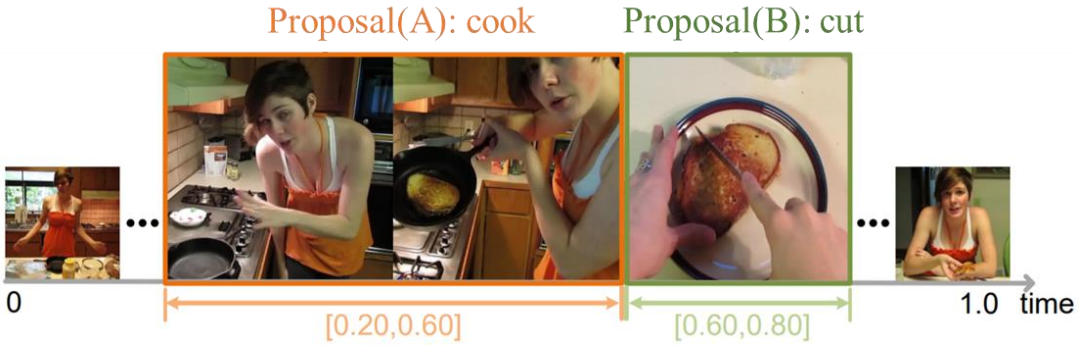
Peijun Bao, Yadong Mu, Learning Sample Importance for Cross-Scenario Video Temporal Grounding, under review

Visual Bias

- Visual concepts queried follow a long-tail distribution
- Some visual concept are much more frequently queried:

  e.g. #stand > #ride (# denotes the frequency)

Model can “predict” the results according to the statistics of visual concepts.

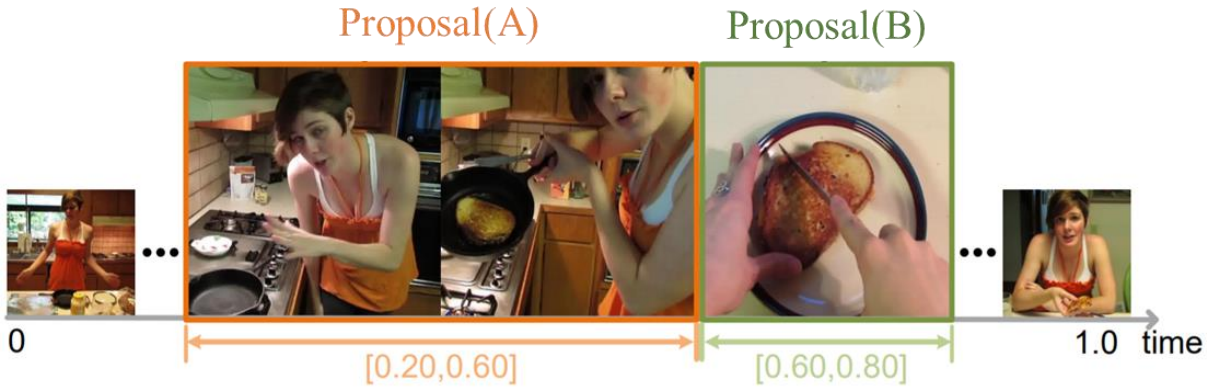


#cook > #cut ⇒ Prior(Proposal A) > Prior(Proposal B)

Temporal Bias

- Temporal intervals also follow a long-tail distribution: e.g. #[0, 1.00] > #[0, 0.19] (temporal intervals are normalized here)

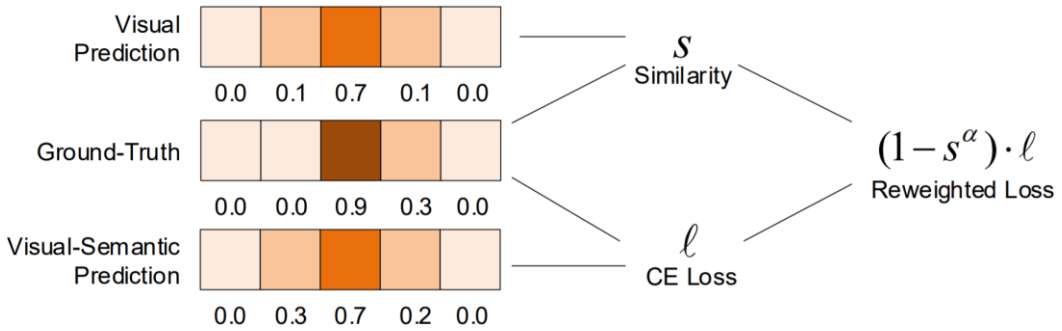
Model can “infer” the results due to the bias of temporal intervals.



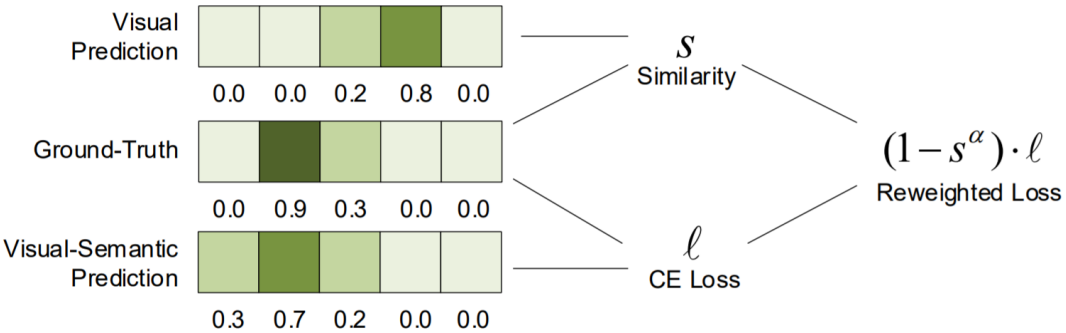
#[0.20,0.60] > #[0.60,0.80] ⇒ Prior(Proposal A) > Prior(Proposal B)

Sample Importance Reweighing

- The visual localizer adjusts the loss function for the visual-semantic localizer according to sample importance.



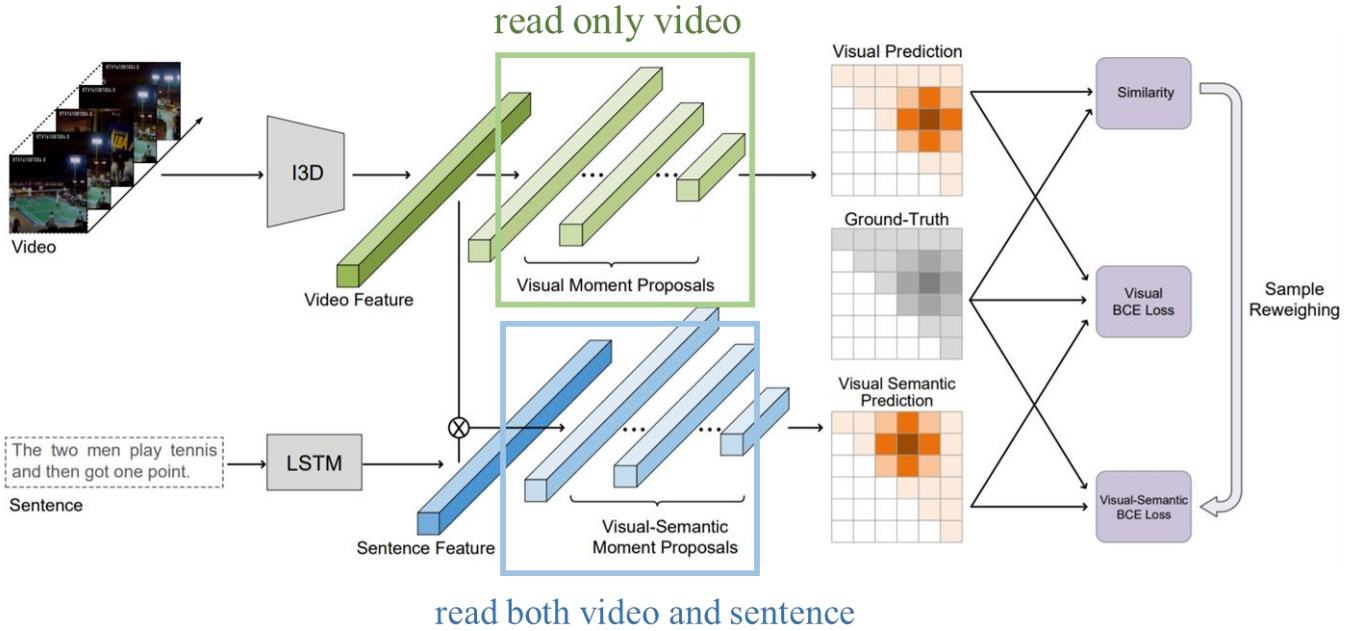
suppress the importance of training sample with high relevance to video moment bias



augment the weight of the irrelevant one

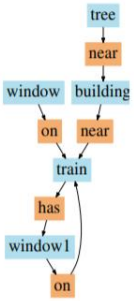
Cross-Scenario Temporal Grounding

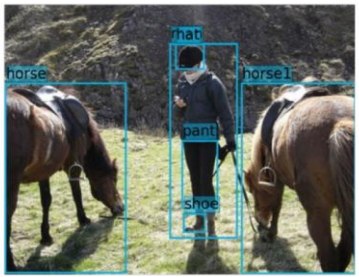
- Simultaneously train two twined models
- One of them aiming to learn video moment bias and further to debias the other model.

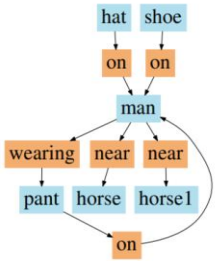


Visual Relation Detection and Scene Graph

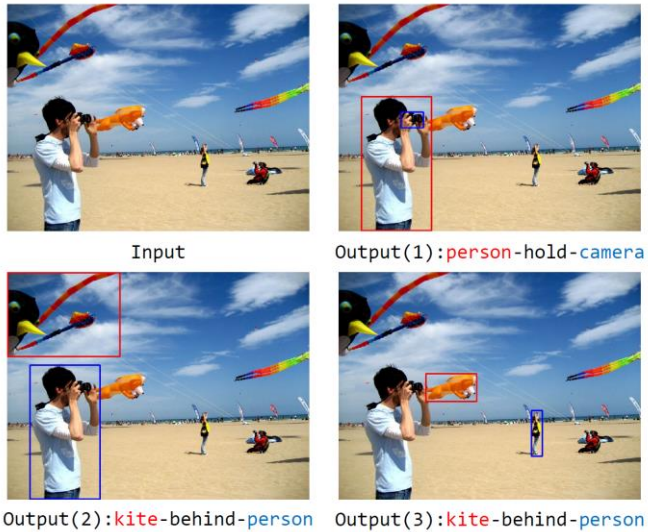
- a graphical data structure: $G = (O, E)$, where $O = (o_1, \ldots, o_n)$ is a set of objects and $E \subseteq O \times R \times O$ is a set of edges.





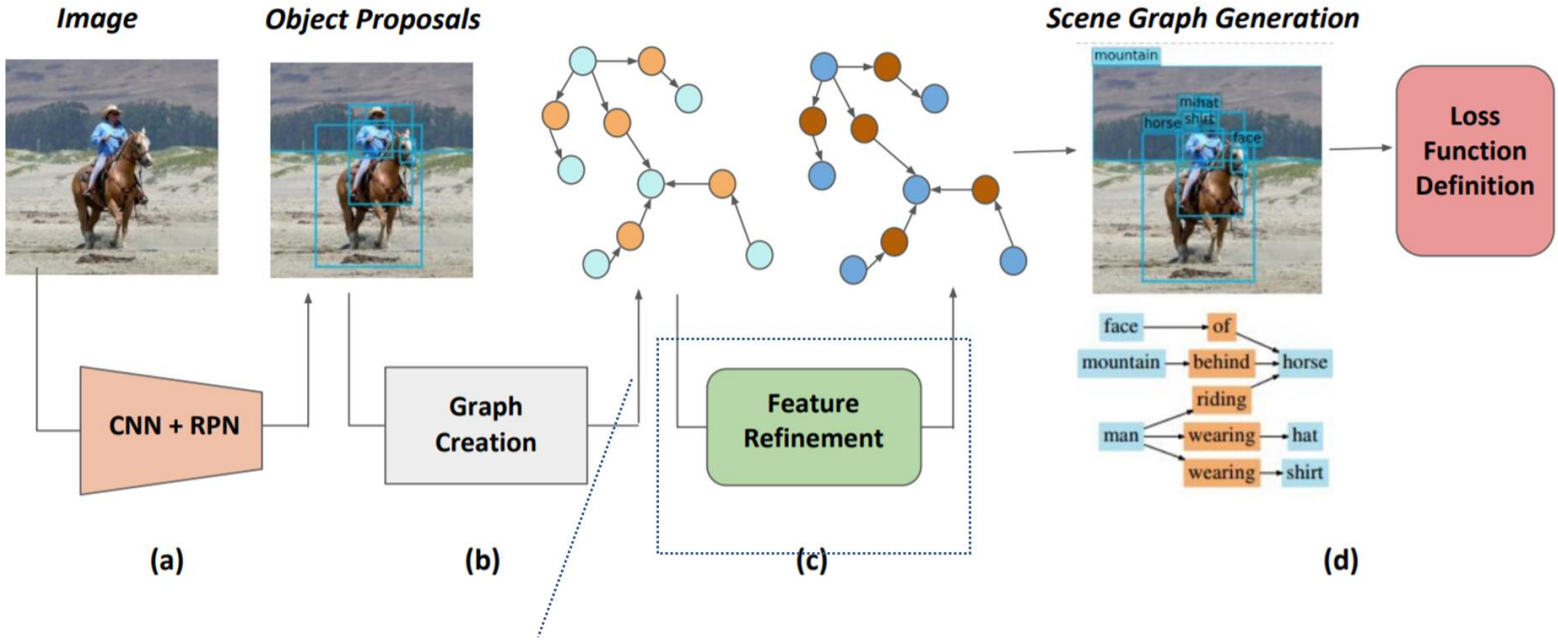


Scene graph



Subject-predicate-object triplet

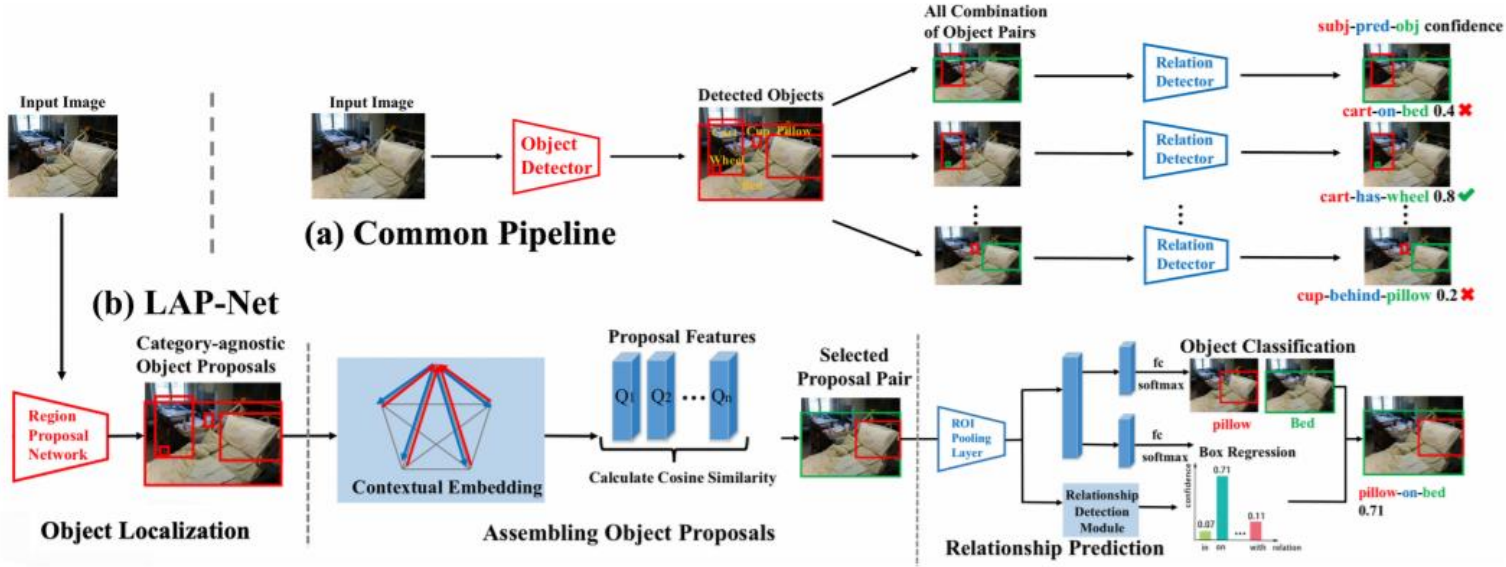
Common Pipeline



Key Module: Message Passing Between Nodes

Localize-assemble-predicate Network (LAP-Net)

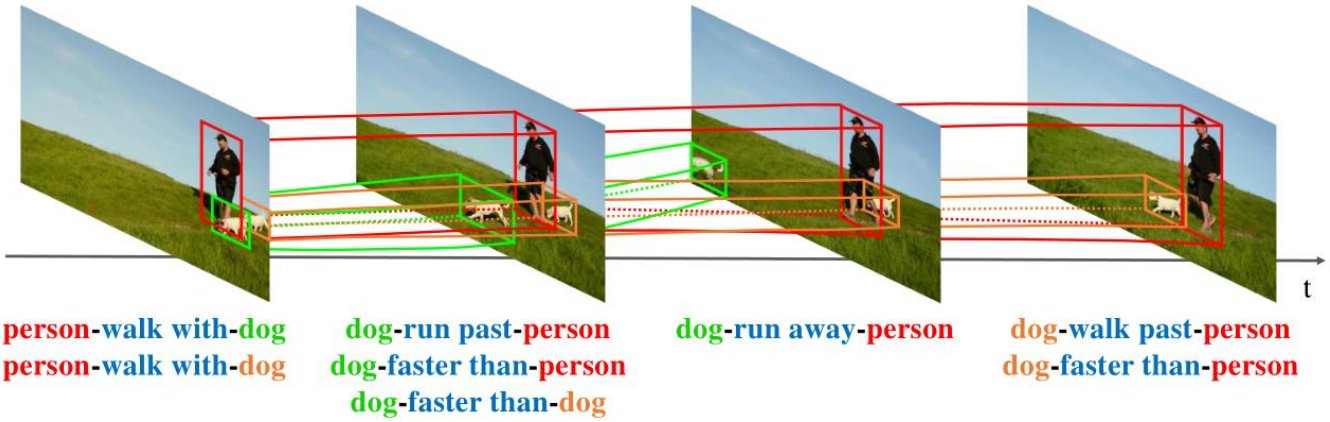
- valid relations combinatorially grow in $O(C^2R)$ for C object categories and R relationships.
- Three stages: : localizing individual objects, assembling and predicting the subject-object pairs



Ruihai Wu, Kehan Xu, Chenchen Liu, Nan Zhuang, Yadong Mu, Localize, Assemble, and Predicate: Contextual Object Proposal Embedding for Visual Relation Detection, AAAI-2020 (Oral Presentation)

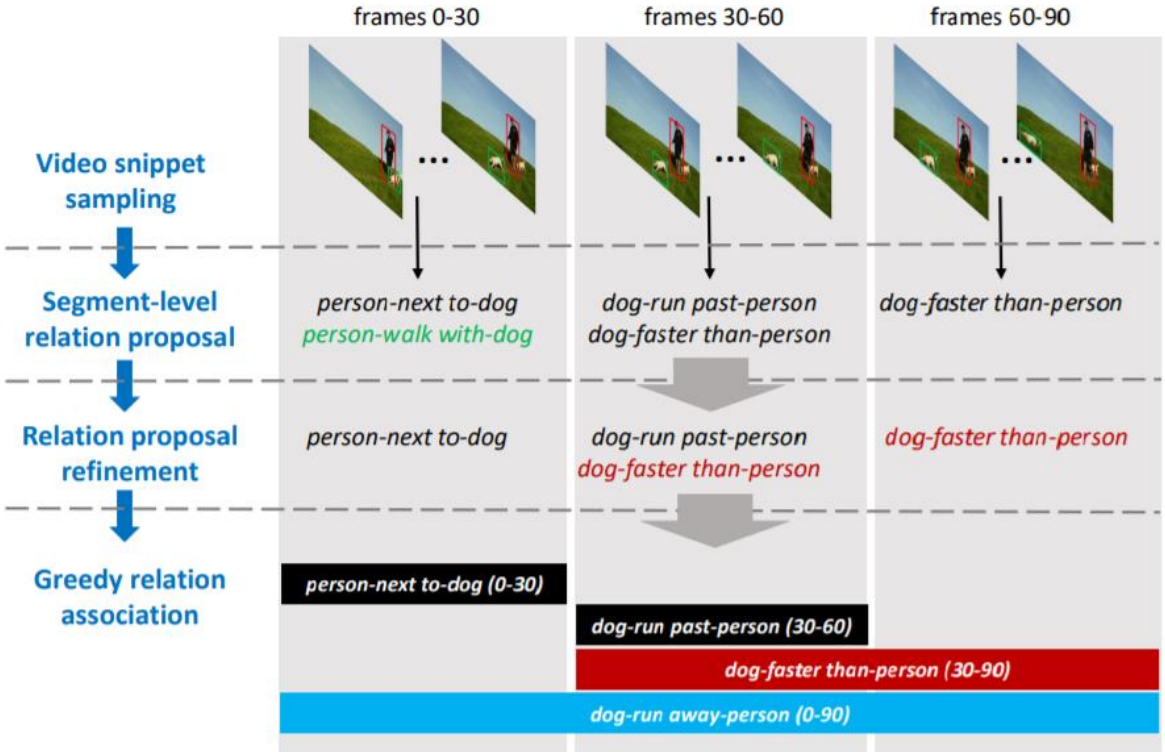
Video Visual Relation Detection

- The visual relation instance is still represented by a relation triplet $\langle subject, predicate, object \rangle \in C \times P \times C$, with the trajectories of the subject and object, F_s and F_o.
- Videos contain dynamic relations like "A-follow-B" and "A-towards-B", and temporally changing relations like "A-chase-B" followed by "A-hold-B"



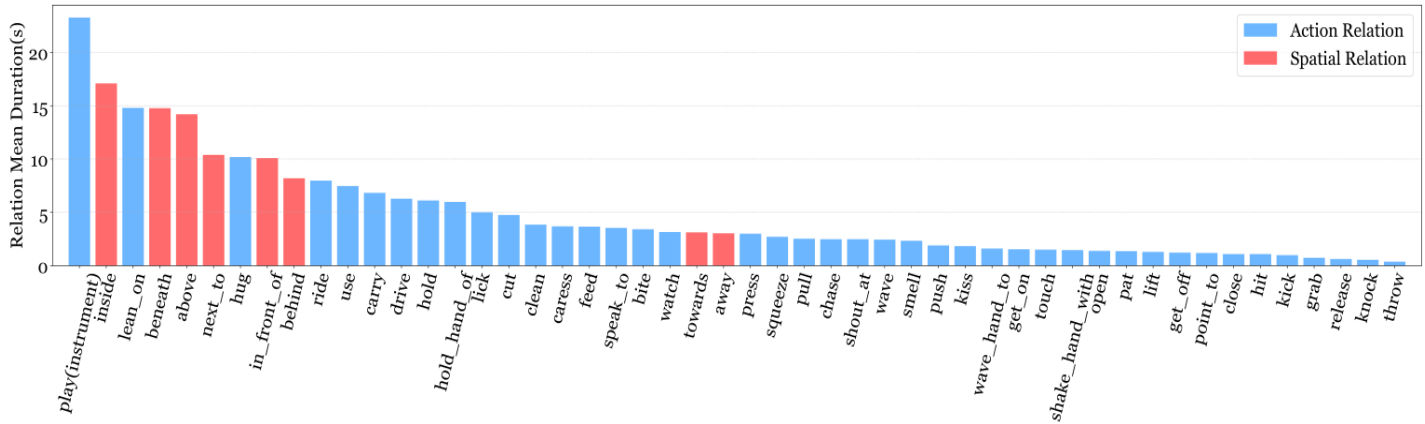
Video Visual Relation Detection

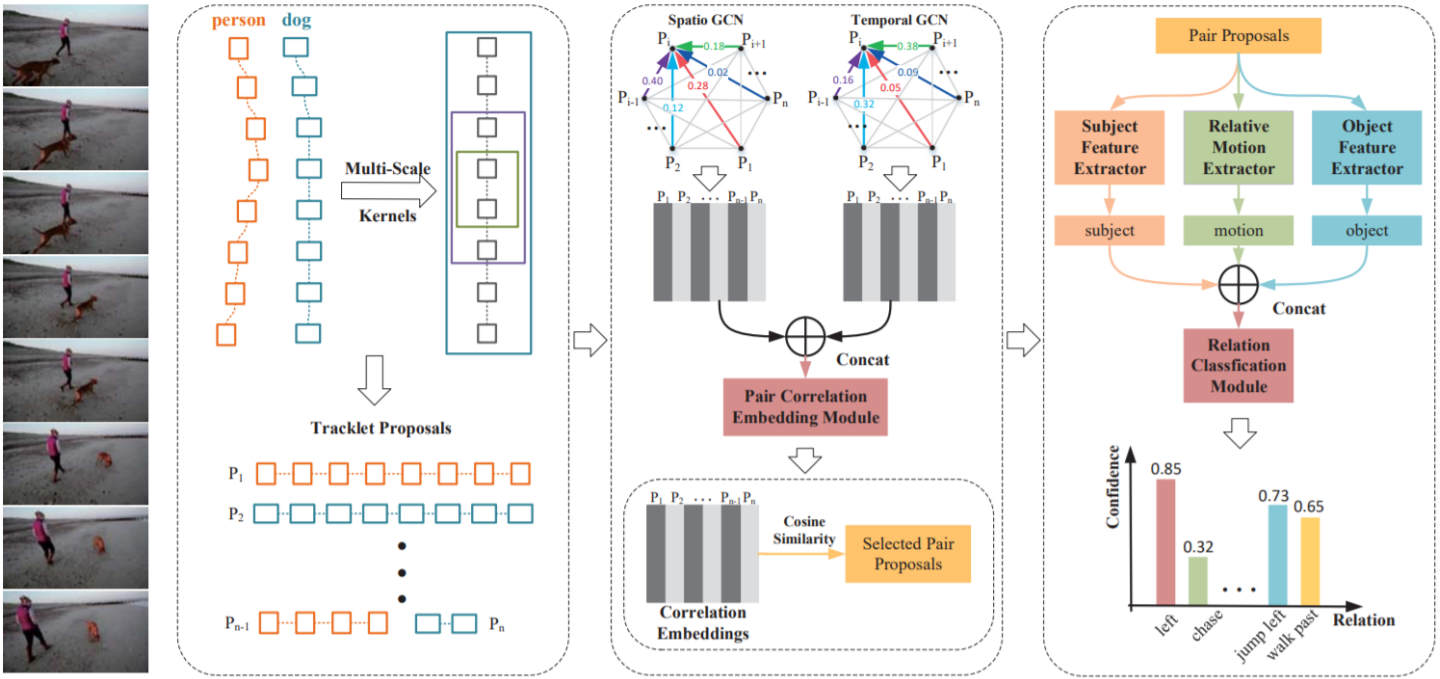
- Common pipeline of existing methods



Video Visual Relation Detection

- Ignoring global information
  - Deal with short video clips, the global Spatio-Temporal information between objects is ignored
- Redundancy in the detection.
  - There is redundancy in the detection of each video clip, which causes the low efficiency.



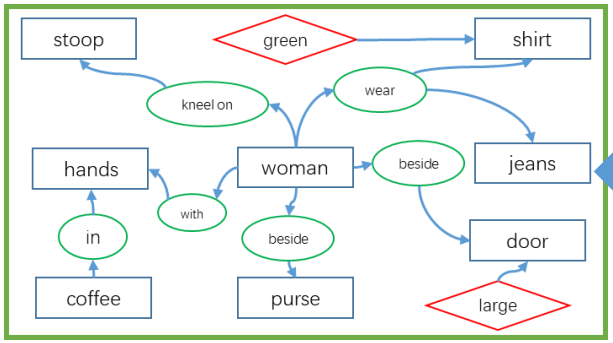

Stage1: Object Tracklets Proposal
Stage2: Relationship Pair Proposal
Stage3: Relationship Classification
person
dog
Multi-Scale
Kernels
Tracklet Proposals
Spatio GCN
Temporal GCN
Concat
Pair Correlation Embedding Module
Cosine Similarity
Selected Pair Proposals
Correlation Embeddings
Pair Proposals
Subject Feature Extractor
Relative Motion Extractor
Object Feature Extractor
subject
motion
object
Concat
Relation Classfication Module
Confidence
0.85
0.32
0.73
0.65
left
chase
jump left
walk past
Relation

High-Order Model for Visual-Text Matching

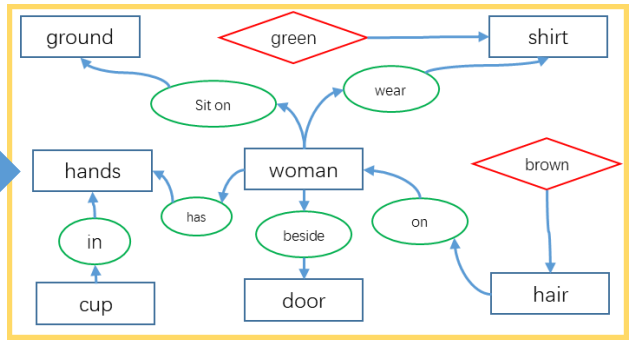
a woman in a green shirt and jeans kneels on a stoop with coffee in hand her purse beside her and a large door

Syntactic analysis

visual object detection



?



Yongzhi Li, Duo Zhang, Yadong Mu, Visual-Semantic Matching by Exploring High-Order Attention and Distraction, IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2020.

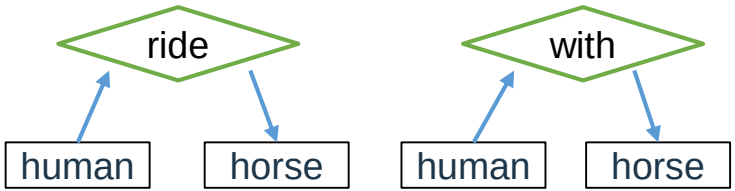
Necessity of High-Order Information

(A) (B) (C)

A man **riding on** a horse on a street

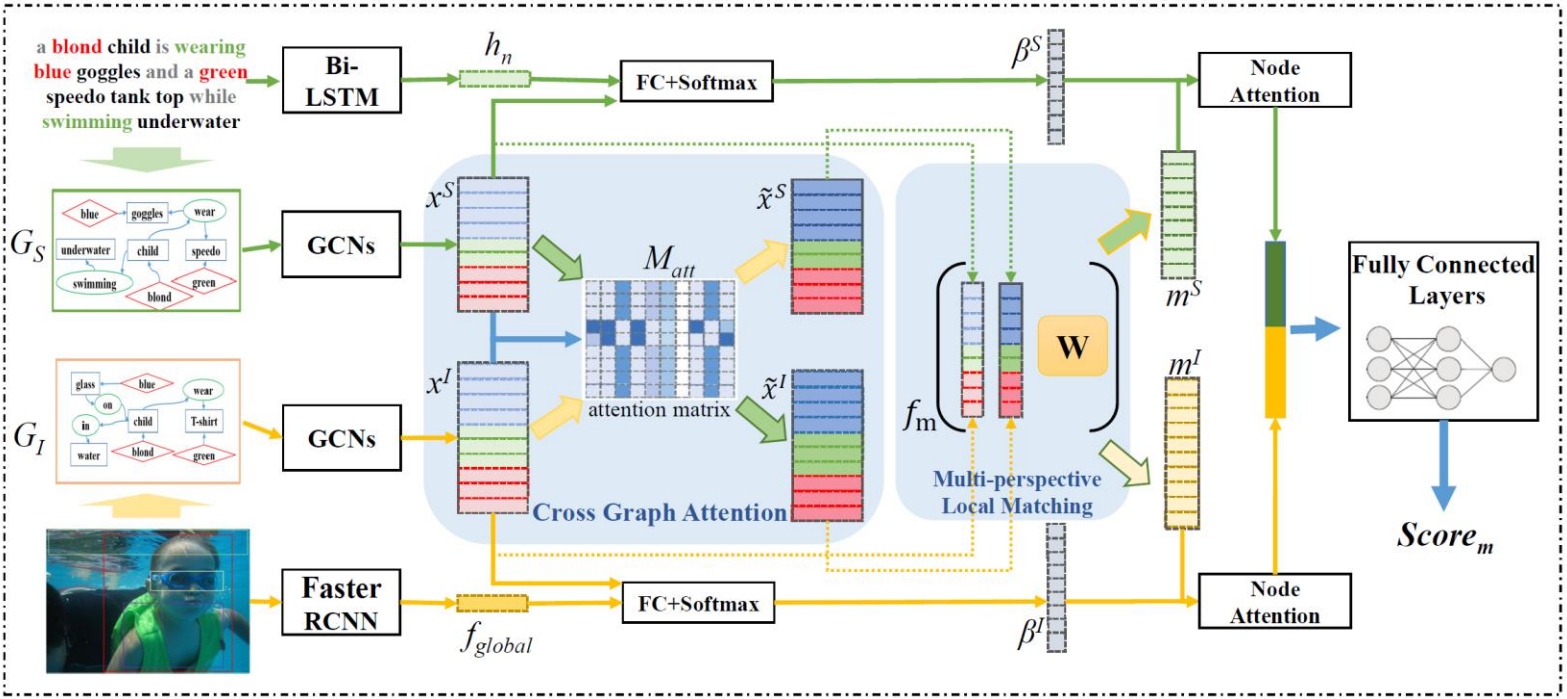
A **brown** dog laying in the grass with a **yellow** frisbee

distinguished by high-order info.



Our Model

- Adopt graph convolution for squeezing and propagating high-order information
- Self-attention + cross-modality attention



Visualization

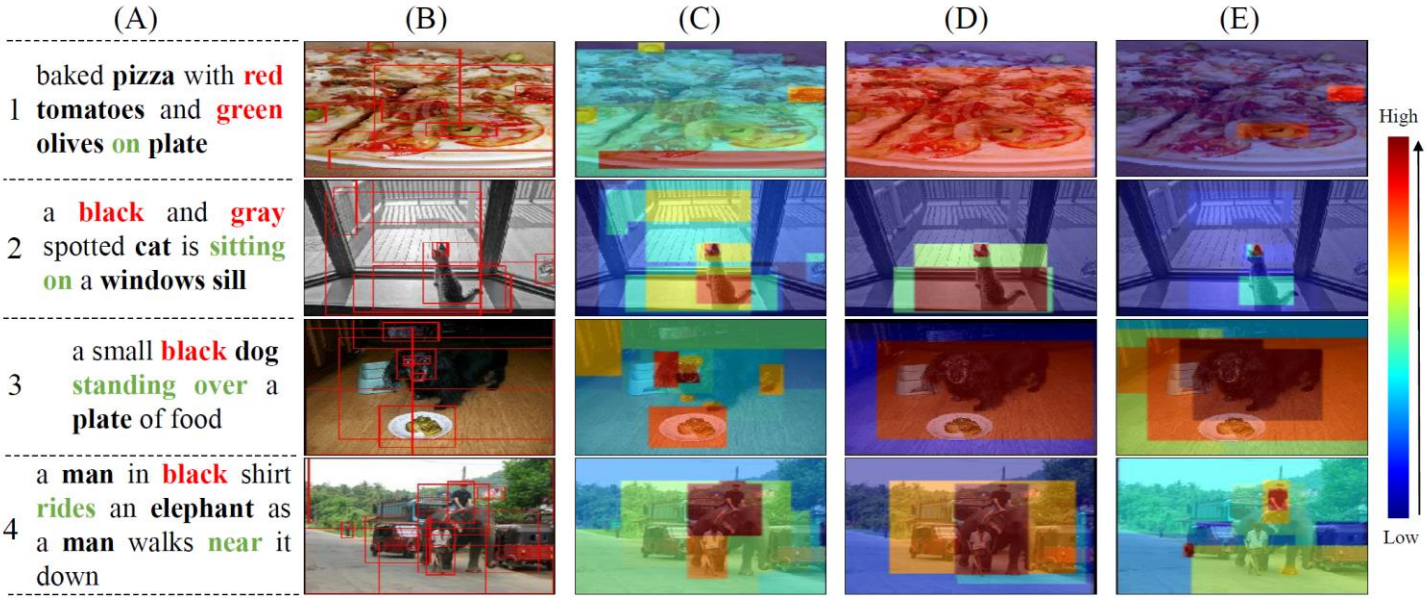
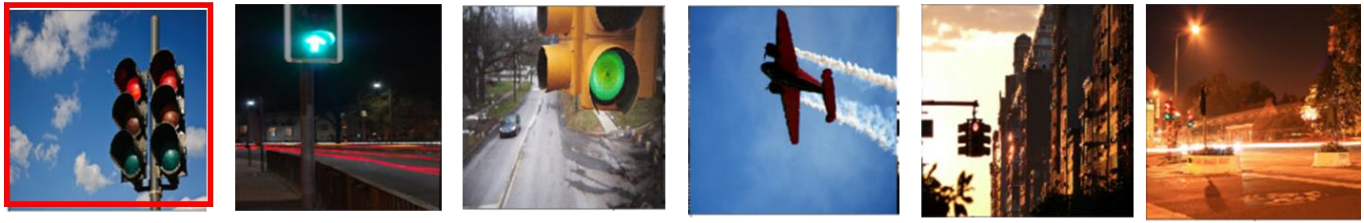


Figure 5. Visualization the cross graph attention mechanism. Each row is an example, and the column (A) shows the query sentences. Part of object proposals are shown in column (B). And columns (C), (D), and (E) show the attention results for the object, relation, and attribute nodes, respectively. The color of the mask reflects the attention value of the area corresponding to the node. The warmer red color represents greater attentive response. Best viewed in color.

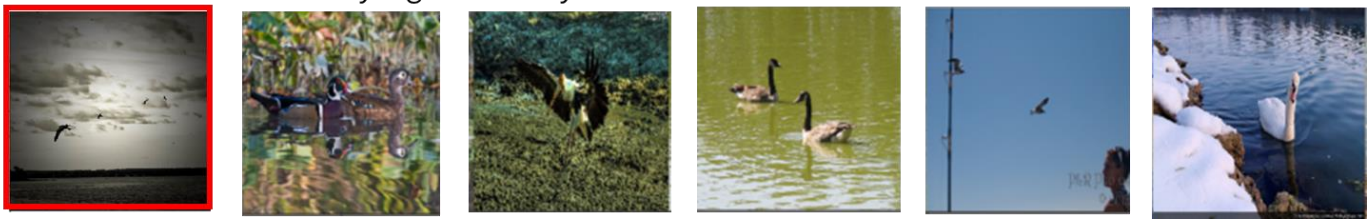
Some Examples

a man with a red helmet on a small moped on a dirt road

a couple of traffic lights sitting under a cloudy sky



a flock of small birds flying in the sky over the water



High-Order Concept Bank for Zero-Shot Search

TRECVID MED 2010 Track Event Definition

Pre-Specified Events

MED '11 Events	*New Events*
Changing a vehicle tire	Attempting a bike trick
Getting a vehicle unstuck	Cleaning an appliance
Grooming an animal	Dog show
Making a sandwich	Giving directions to a location
Parkour	Marriage proposal
Repairing an appliance	Renovating a home
Working on a sewing project	Rock climbing
Birthday party	Town hall meeting
Flash mob gathering	Winning a race without a vehicle
Parade	Working on a metal crafts project

Ad Hoc Events

New Events

Doing homework or studying
Hide and seek
Hiking
Installing flooring
Writing text

Feeding an animal

Landing a fish

Wedding ceremony

High-Order Concept Bank for Zero-Shot Search

Event name: Bee keeping
Definition: One or more people perform activities associated with the keeping of honeybees.
Explication: Bee keeping refers to the maintenance of honeybees by humans. A beekeeper keeps bees in order to collect products of the hive to pollinate crops, or to produce bees for sale.
Evidence: bee, bee keeper, smoke, honey, knife.

新的多媒体事件的文字描述

+

0个该事件的正样本实例

视频语义概念（semantic concepts)

000 - Parade
Name: Parade **Labeled**: Yes
Definition: Multiple units of marchers, devices, bands, banners or Music.

001 - Exiting_Car
Name: Exiting_Car **Labeled**: Yes
Definition: A car exiting from somewhere, such as a highway, building, or parking lot.

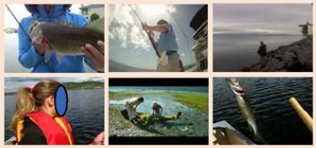
002 - Handshaking
Name: Handshaking **Labeled**: Yes
Definition: Two people shaking hands. Does not include hugging or holding hands.

...

+

基于标注数据的多媒体事件检测模型

Wedding ceremony



Landing a fish

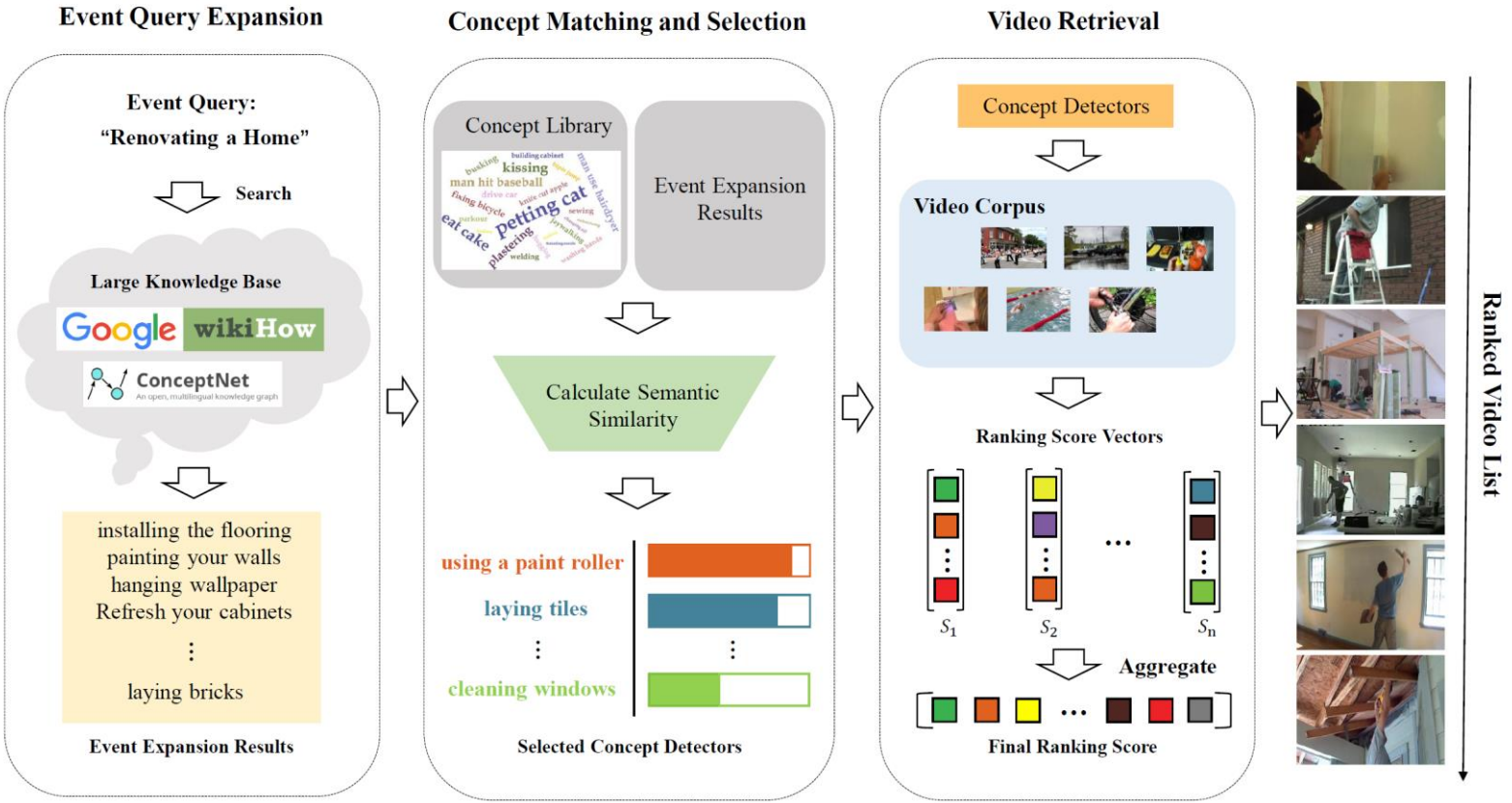
TRECVID MED16
~260000视频片段

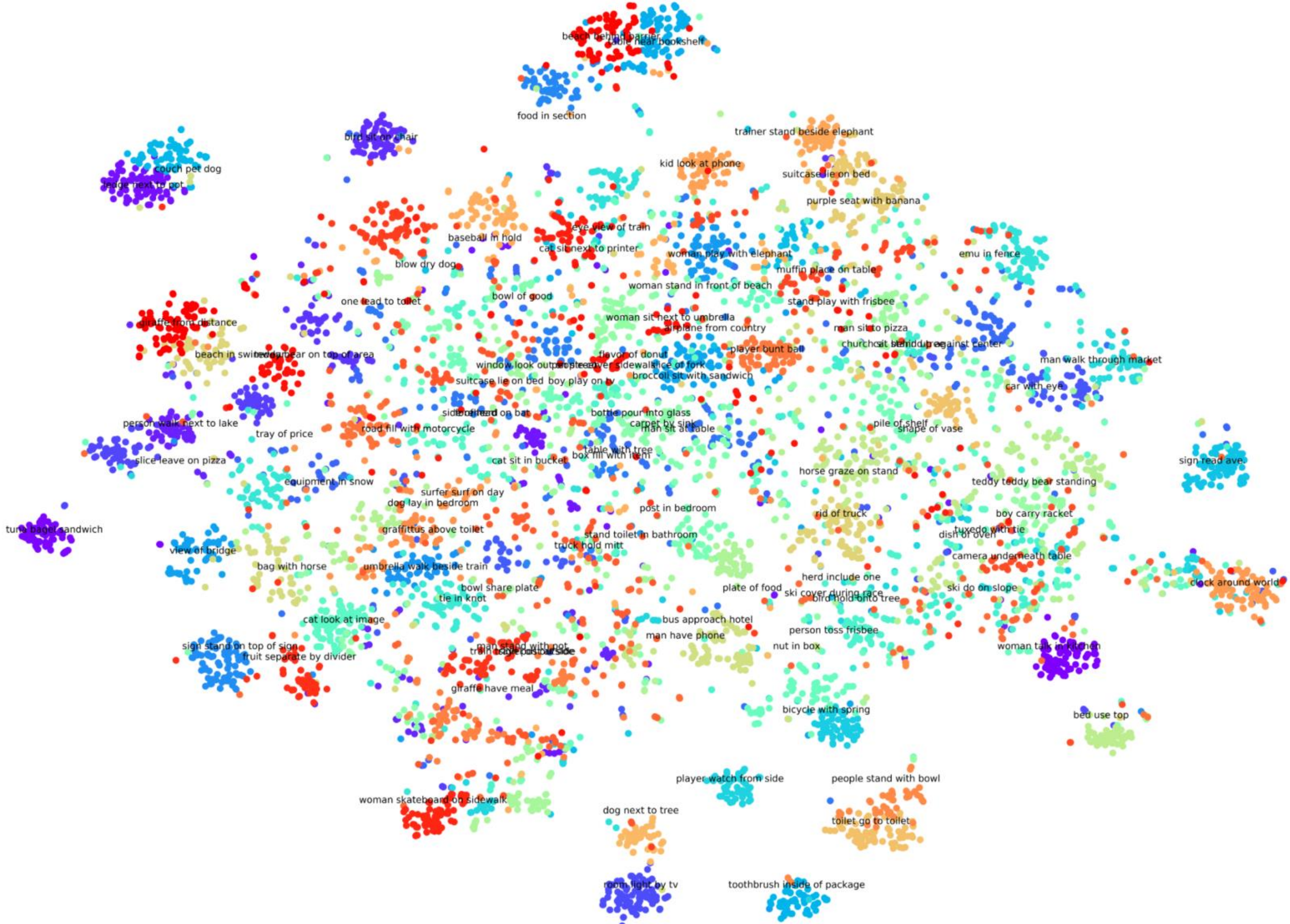
零样本学习

- 语义概念选择
- 语义事件的流形结构
- 排序融合（rank aggregation）

零样本学习的检测结果

High-Order Concept Bank




beach behind barrier
table near bookshelf
food in section
trainer stand beside elephant
bird sit on chair
couch pet dog
ledge next to pot
kid look at phone
suitcase lie on bed
purple seat with banana
eye view of train
baseball in hold
cat sit next to printer
woman play with elephant
emu in fence
blow dry dog
muffin place on table
woman stand in front of beach
bowl of good
one lead to toilet
stand play with frisbee
woman sit next to umbrella
giraffe from distance
airplane from country
man sit to pizza
player bunt ball
flavor of donut
beach in swim
man walk through market
window look out
broccoli sit with sandwich
suitcase lie on bed
boy play on tv
car with eye
bottle pour into glass
person walk next to lake
road fill with motorcycle
pile of shelf
shape of vase
tray of price
man sit at table
table with tree
box fill with item
cat sit in bucket
slice leave on pizza
sign read ave.
horse graze on stand
equipment in snow
teddy teddy bear standing
surfer surf on day
dog lay in bedroom
post in bedroom
rid of truck
boy carry racket
tuna bagel sandwich
graffittus above toilet
tuxedo with tie
dish of oven
stand toilet in bathroom
truck hold mitt
view of bridge
camera underneath table
bag with horse
umbrella walk beside train
herd include one
bowl share plate
plate of food
clock around world
ski do on slope
tie in knot
bird hold onto tree
cat look at image
bus approach hotel
person toss frisbee
man have phone
sign stand on top of sign
nut in box
woman talk in kitchen
fruit separate by divider
giraffe have meal
bicycle with spring
bed use top
player watch from side
people stand with bowl
woman skateboard on sidewalk
dog next to tree
toilet go to toilet
room light by tv
toothbrush inside of package

Birthday Party

blowing out candles, children eat cake, celebrating, popping balloons

Flash Mob Gathering

dancing macarena, square dancing, group of people on street

Grooming an Animal

bathing dog, petting cat, cutting nails, woman hold a dog

Working On a Metal Crafts Project

welding, bending metal, making horseshoes, using a wrench

Getting a Vehicle Unstuck

wading through mud, driving car, shoveling snow, people push car

Tuning Musical Instrument

playing guitar, tapping guitar, playing organ, playing keyboard

Assembling Bicycle

assembling bicycle, fixing bicycle, man stand with bicycle, riding a bike

Getting a Haircut

combing hair, curling hair, fixing hair, washing hair, woman dry hair

Cleaning Sink

cleaning toilet, washing dishes, washing hands, sink in bathroom

Making a Cake

making a cake, eating cake, cake on plate, piece of cake, woman in kitchen

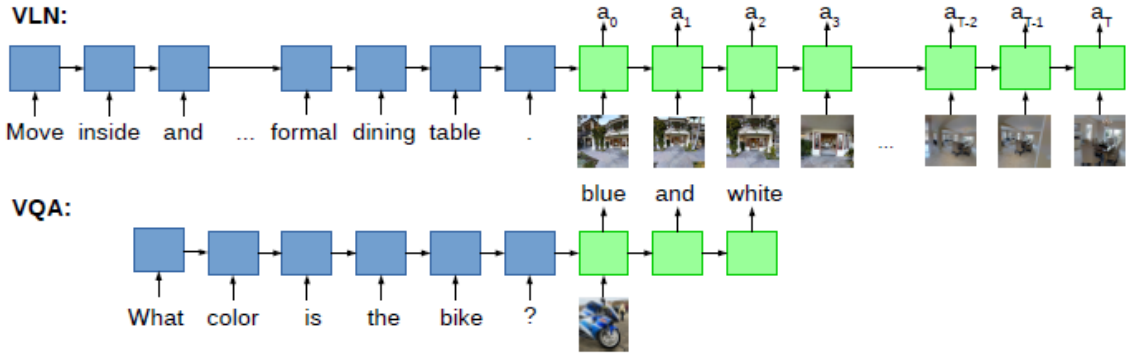
Vision-and-Language Navigation (VLN)

- A new trending topic in vision-language field featuring combining CV, NLP, and robotic control.
- Task formulation:
  - An embodied agent positioned in a 3D environment
  - Given a natural language instruction describing a route in the environment, the agent is required to follow the instruction and navigate to the target location autonomously

Leave the bedroom, and enter the kitchen. Walk forward, and take a left at the couch. Stop in front of the window.

Vision-and-Language Navigation (VLN)

- A basic framework: seq-to-seq model



	Trajectory Length (m)	Navigation Error (m)	Success (%)	Oracle Success (%)
Test (unseen):				
SHORTEST	9.93	0.00	100	100
RANDOM	9.93	9.77	13.2	18.3
Human	11.90	1.61	86.4	90.2
Student-forcing	8.13	7.85	20.4	26.6

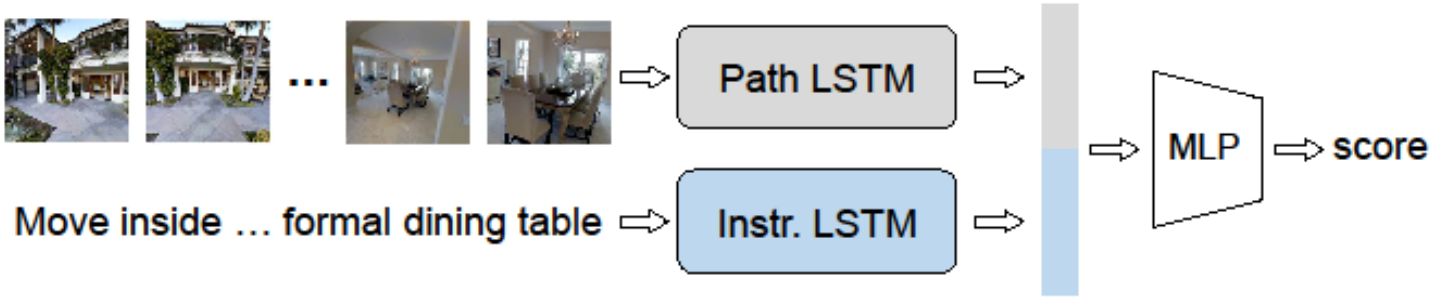
Spatial Route Prior

- Rethinking the dominating RL-based solutions: routes are not random, so are the instructions
- Shortest paths are common

Xinzhe Zhou, Wei Liu, Yadong Mu, Rethinking the Spatial Route Prior in Vision-and-Language Navigation, under review

The Naïve case: Known Map+ Shortest Path

- Under shortest-path prior (e.g., R2R)
  - Every position is uniquely associated with one shortest-path from the starting point (ignore the case of multiple shortest-paths)
  - If the topology is known, the path-finding problem could be equivalently transformed into a position / node classification problem



Relaxation: Known Map + General Route Prior

- Every route could be decomposed into several connected shortest-paths
- Sequential decision among the shortest paths instead of the instant neighbors

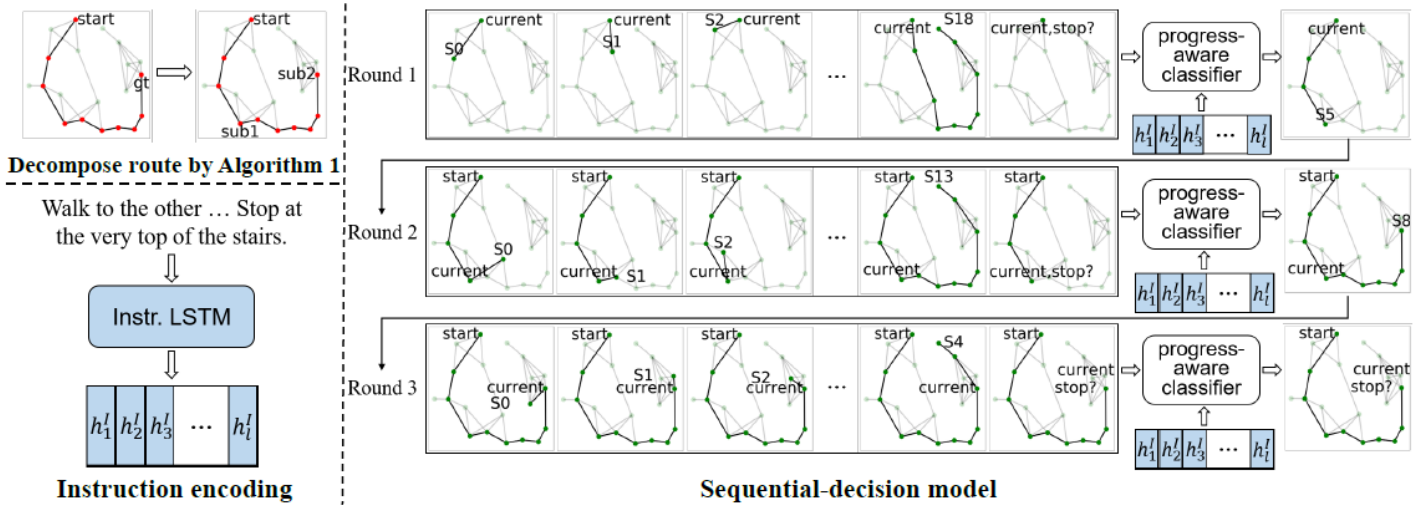


Figure 1: An example illustrating the route decomposition and the sequential-decision process.

Relaxation: Unknown Map + General Route Prior

- w/o map: pre-exploration + path ranking

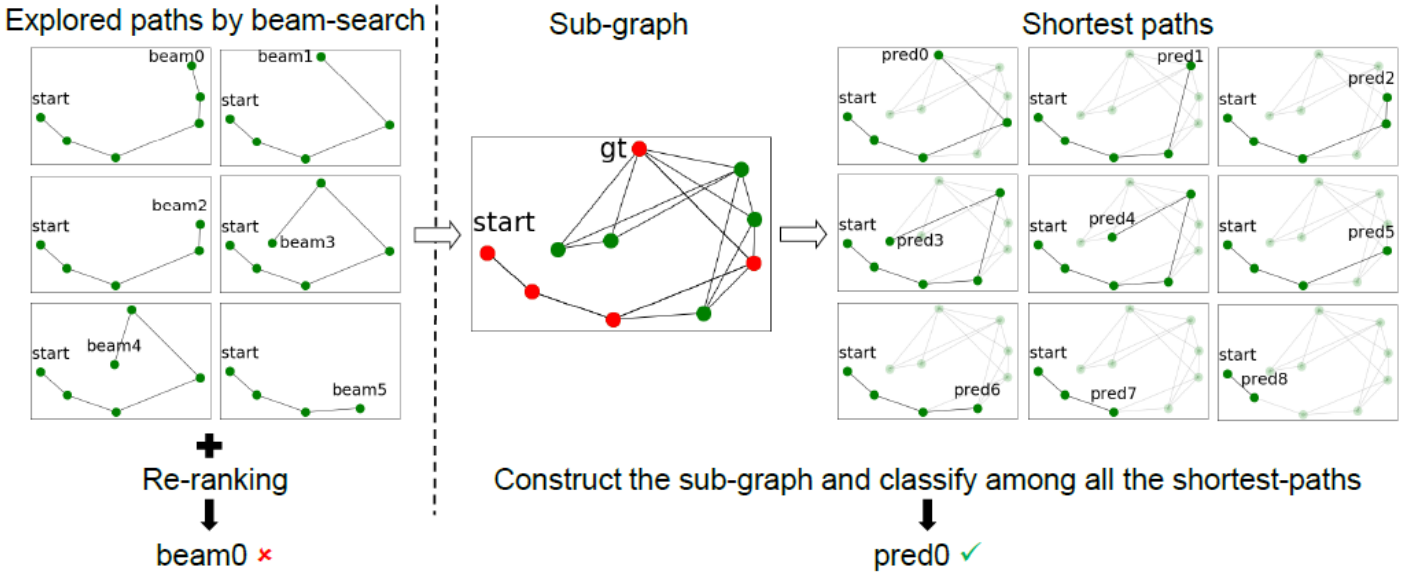


Figure 2: An example illustrating our explore-and-exploit scheme (right) with comparison to previous beam-search and re-ranking paradigm (left) from R2R.

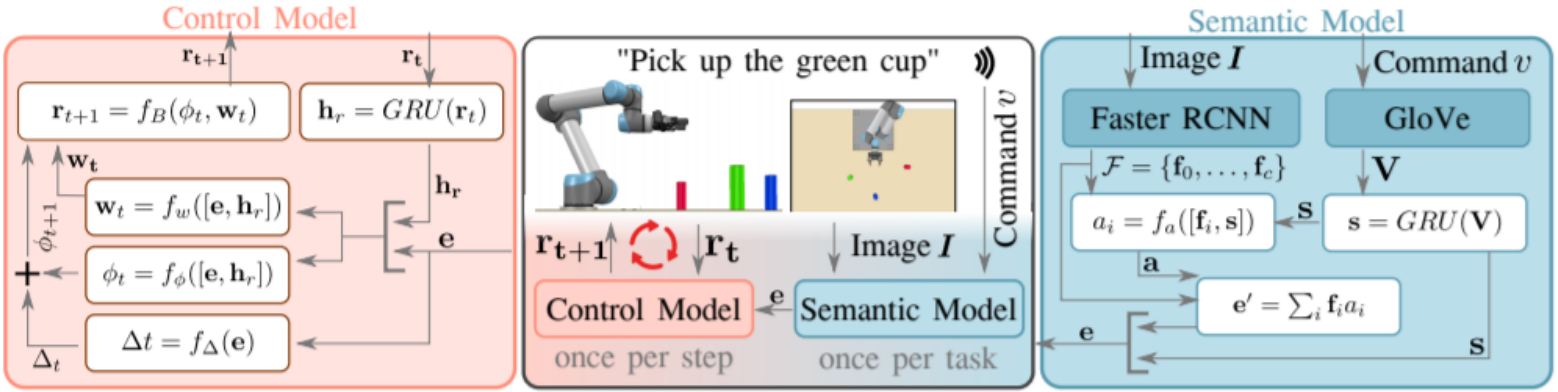
Experimental Results on R2R and R4R (and RxR)

Models	R2R														
	val-seen					val-unseen					test-unseen				
	SR↑	NE↓	TL↓	OSR↑	SPL↑	**SR↑**	NE↓	TL↓	OSR↑	SPL↑	**SR↑**	NE↓	TL↓	OSR↑	SPL↑
single-run															
Speaker-Follower [8]	0.66	3.36	-	0.74	-	0.36	6.62	-	0.45	-	0.35	6.62	14.8	0.44	0.28
RCM [29]	0.67	3.53	10.7	0.75	-	0.43	6.09	11.5	0.50	-	0.43	6.12	12.0	0.50	0.38
Self-Monitoring [20]	0.67	3.22	-	0.78	0.58	0.45	5.52	-	0.56	0.32	0.43	5.99	18.0	0.55	0.32
Tactical Rewind [16]	-	-	-	-	-	0.56	4.97	21.2	-	0.43	0.54	5.14	22.1	0.64	0.41
EnvDrop [25]	0.62	3.99	11.0	-	0.59	0.52	5.22	10.7	-	0.48	0.51	5.23	11.7	0.59	0.47
AuxRN [32]	0.70	3.33	-	0.78	0.67	0.55	5.28	-	0.62	0.50	0.55	5.15	-	0.62	0.51
Active Perception [27]	0.70	3.20	19.7	**0.80**	0.52	0.58	4.36	20.6	0.70	0.40	0.60	4.33	21.6	0.71	0.41
VLN-BERT [22]	0.70	3.73	10.3	0.76	0.66	0.59	4.10	9.6	0.69	0.55	-	-	-	-	-
SSM [26]	0.71	**3.10**	14.7	**0.80**	0.62	0.62	4.32	20.7	0.73	0.45	0.61	4.57	20.4	0.70	0.46
Ours(known map)	**0.76**	3.69	9.7	0.79	**0.72**	**0.74**	**3.56**	9.1	**0.80**	**0.71**	**0.72**	**3.77**	9.4	**0.76**	**0.69**
Ours(explore@40)	0.74	3.72	**9.6**	0.77	0.71	0.73	3.61	**9.0**	0.78	0.70	0.71	3.81	**9.3**	0.75	**0.69**

Models	R4R											
	val-seen						val-unseen					
	SR↑	NE↓	TL	**CLS↑**	nDTW↑	SDTW↑	SR↑	NE↓	TL	**CLS↑**	nDTW↑	SDTW↑
Speaker-Follower [8]	0.52	5.35	15.4	0.46	-	-	0.24	8.47	19.9	0.30	-	-
RCM(goal oriented) [15]	0.56	5.11	24.5	0.40	-	-	0.29	8.45	32.5	0.20	0.27	0.11
RCM(fidelity oriented) [15]	0.53	5.37	18.8	0.55	-	-	0.26	8.08	28.5	0.35	0.30	0.13
PTA(low-level) [17]	0.57	5.11	11.9	0.52	0.42	0.29	0.27	8.19	10.2	0.35	0.20	0.08
PTA(high-level) [17]	0.58	**4.54**	16.5	0.60	**0.58**	0.41	0.24	8.25	17.7	0.37	0.32	0.10
EGP [6]	-	-	-	-	-	-	0.30	8.00	18.3	0.44	0.37	0.18
EnvDrop [25]	0.52	-	19.9	0.53	-	0.27	0.29	-	27.0	0.34	-	0.09
OAAM [24]	0.56	-	11.8	0.54	-	0.32	0.31	-	13.8	0.40	-	0.11
SSM [26]	**0.63**	4.60	19.4	0.65	0.56	**0.44**	0.32	8.27	22.1	0.53	0.39	0.19
Ours(known map)	0.45	6.92	20.1	**0.67**	0.57	0.38	**0.36**	7.46	20.5	**0.61**	**0.50**	**0.28**
Ours(explore@40)	0.45	6.35	19.8	0.66	**0.58**	0.36	**0.36**	**7.02**	20.2	0.60	**0.50**	0.27

The Next: Language-Instructed Embodied AI

- Vision + Language + Control (Robotics)
- Key functionalities: grounding / reasoning / 3-D estimation



Questions?

Email: myd@pku.edu.cn
Website: http://www.muyadong.com