Multimodal video representations
&
their extension to robot navigation

Cordelia Schmid

Inria / Google

Automatic video understanding

Huge amount of video is available and growing daily

BBC Motion Gallery

ina

TV-channels recorded since 60's

YouTube
Broadcast Yourself

> 500k hours of videos uploaded every minute

Over one billion surveillance cameras world-wide

Why multimodal video representation?

- Precise understanding of the video content
  → Requires access to all modalities simultaneously

Is this Indian?

Why multimodal video representation?

- Large-scale cross-modal supervision
  → No manual annotation required

[HowTo100M. A. Miech, D. Zhukov, JB Alayrac, M. Tapaswi, I. Laptev and J. Sivic, ICCV 2019]

Overview

- ***VideoBERT***
- Dense Video Captioning
- Retrieval Augmented Visual Question Answering
- Multimodal transformer for navigation and manipulation

VideoBERT: Learning from multimodal video

- Learning from visual video and speech transcribed with ASR

But in the meantime, you're just kind of moving around your cake board.

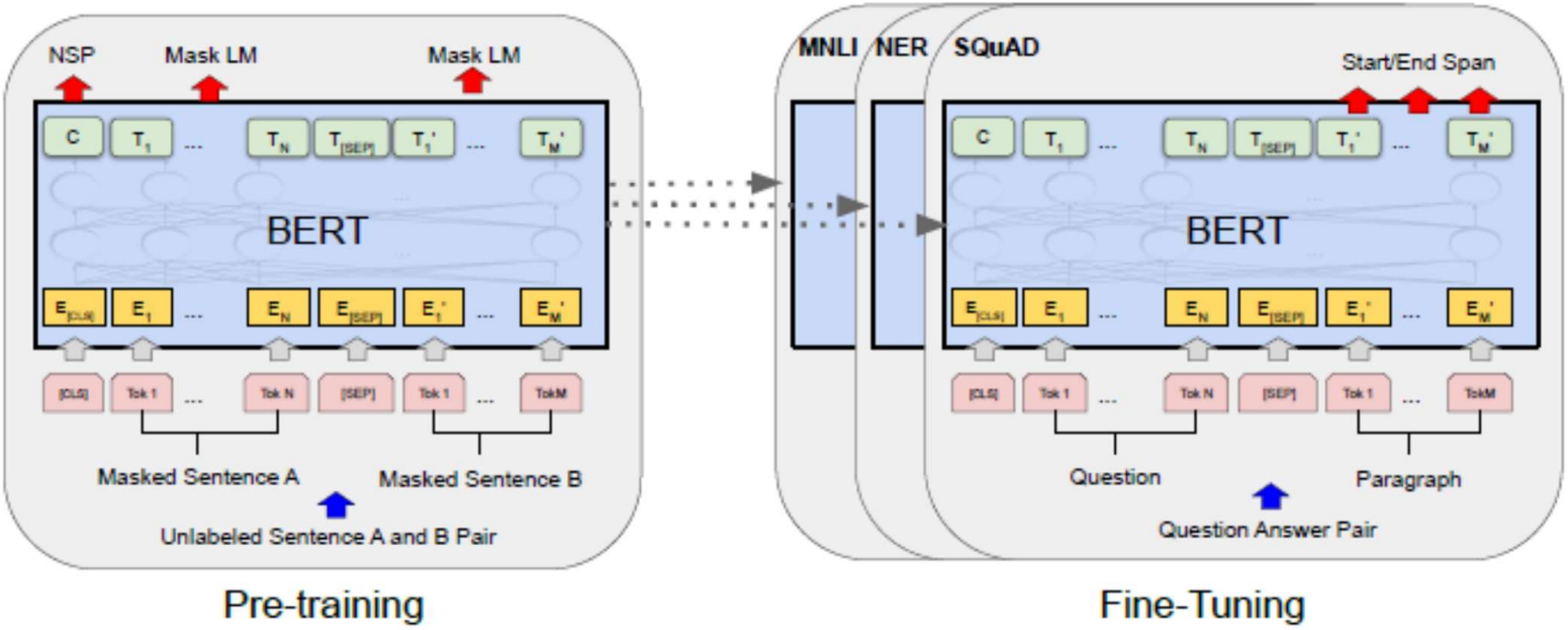
- BERT model learns correspondence between video and speech

- Learning from large-scale data, i.e., cooking, instruction videos

[VideoBERT, C. Sun, A. Myers, C. Vondrick, K. Murphy, C. Schmid, ICCV'19]

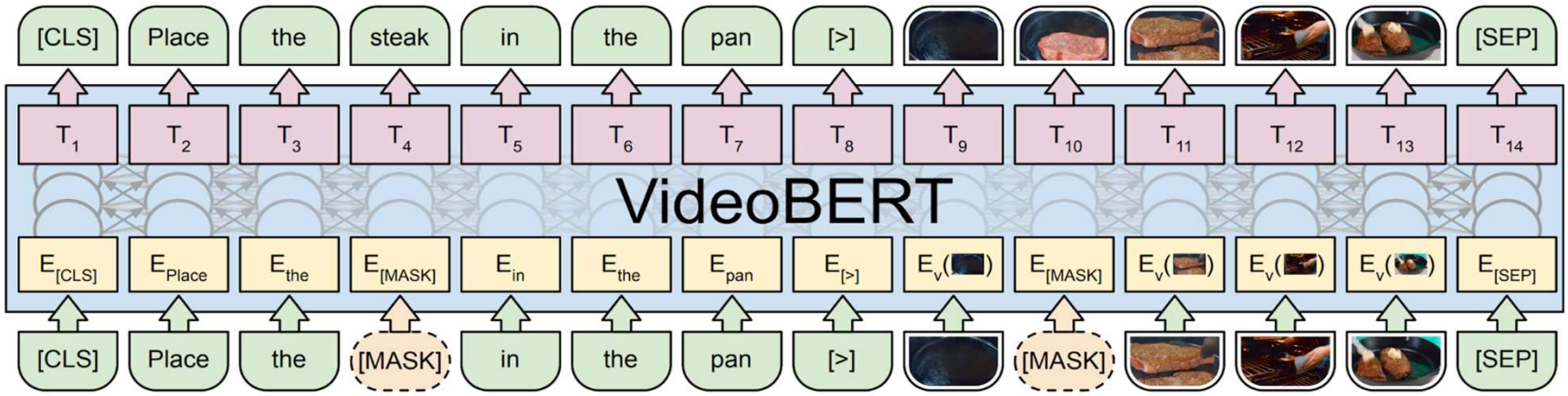
BERT model

BERT: Bidirectional Encoder Representation from Transformers



[BERT, J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, NAACL 2019]

VideoBERT



Text (ASR)

Video (3D-conv features)

- BERT for multiple modalities
- Masked ‘language’ modeling as in BERT
- Video representation with 3D-convolutions + clustering

[VideoBERT, C. Sun, A. Myers, C. Vondrick, K. Murphy, C. Schmid, ICCV’19]

VideoBERT

Training on 300k cooking videos

“Keep rolling tight and squeeze the air out to its side”

Zero-shot prediction

Verb: make, **Noun:** pizza

[VideoBERT, C. Sun, A. Myers, C. Vondrick, K. Murphy, C. Schmid, ICCV 2019]

Zero-shot prediction

Method	Verb (top-5 %)	Object (top-5 %)
S3D (supervised)	**46.9**	30.9
VideoBERT	43.3	**33.7**

Results on YouCook II dataset

Pre-training size	Verb (top-5 %)	Object (top-5 %)
10K	15.5	17.8
50K	15.7	27.3
100K	24.5	30.6
300K	**43.3**	**33.7**

- VideoBERT learns video-language correspondence
- Close to fully-supervised accuracy
- More data improves the performance (not saturated yet)

[VideoBERT, C. Sun, A. Myers, C. Vondrick, K. Murphy, C. Schmid, ICCV 2019]

Fine-tuning on downstream tasks

- For captioning cooking video on YouCook2

Method	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr
Zhou et al. (CVPR'18)	-	1.42	11.20	-	-
S3D	6.12	3.24	10.00	26.05	0.35
VideoBERT	6.80	4.07	10.99	27.51	0.50

- Effective and outperforms S3D features
- Pre-training helps!

[VideoBERT, C. Sun, A. Myers, C. Vondrick, K. Murphy, C. Schmid, ICCV 2019]

Conclusion

- VideoBERT allows to model jointly video + text
- Can be used for zero-shot prediction and pretraining
- Allows to model long-term temporal dependencies

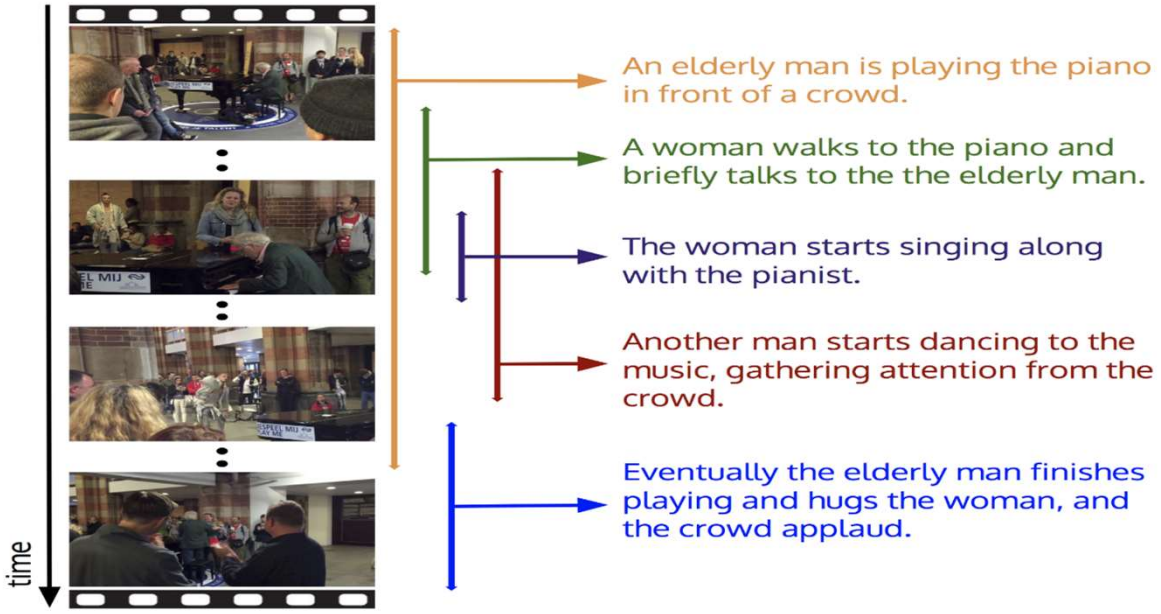
Overview

- VideoBERT
- ***Dense Video Captioning***
- Retrieval Augmented Visual Question Answering
- Multimodal transformer for navigation and manipulation

Dense video captioning - task

Video captioning models for long videos with multiple events

- Captions are grounded in the video
- Combines localization and text generation



Example of dense, overlapping captions from the ActivityNet dataset

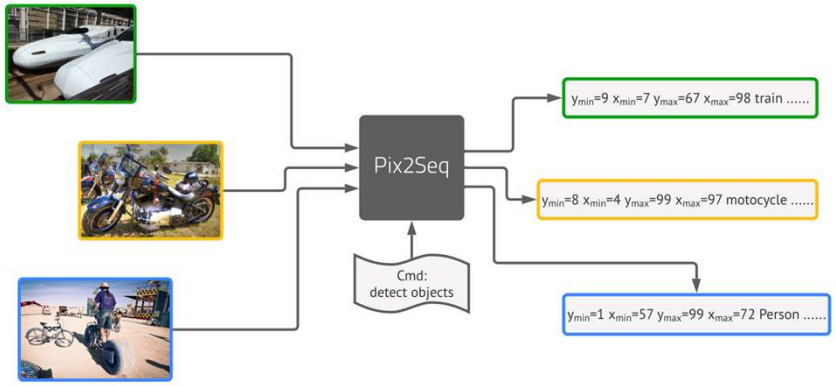
Dense video captioning – SOTA

Current approaches for dense video captioning

- Train separate networks for localization and captioning
- Require task-specific components like event counters
- Train on manually annotated datasets (small)
- Cannot reason over *long* videos

Localization as language modeling

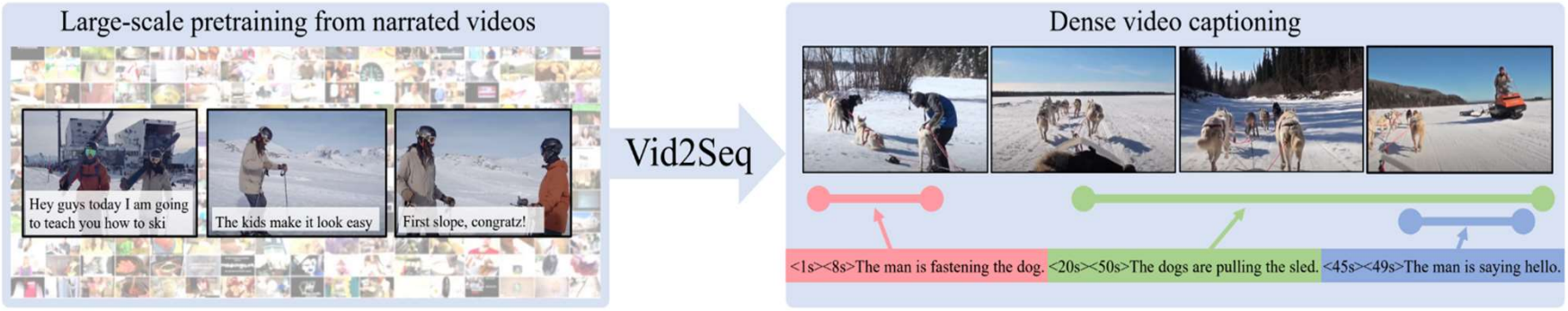
- Pix2seq casts object detection as sequence generation
- Spatial coordinates are quantized and tokenized



Our approach Vid2Seq

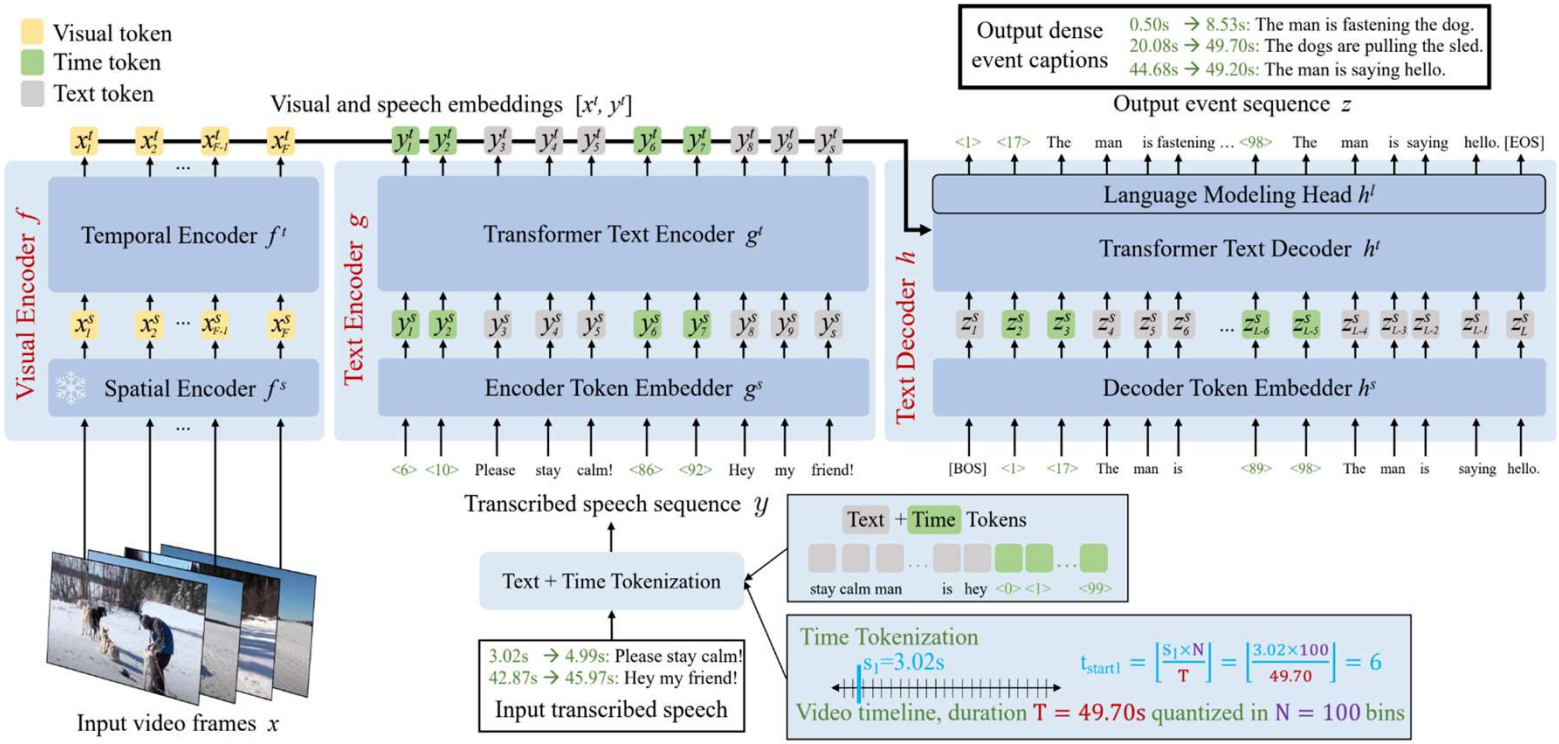
- Single target sequence consists of **Text + Time tokens combining localization + captioning**

- Large-scale pretraining from narrated untrimmed videos



[Vid2Seq, A. Yang, A. Nagrani, P. Seo, A. Miech, J. Pont-Tuset, I. Laptev, J. Sivic, C. Schmid, CVPR 2023]

Vid2Seq – model



- Frozen Visual backbone (CLIP)
- Temporal Encoder for video
- Speech is cast as a single sequence of text and time tokens
- T5 Encoder & Decoder

[Vid2Seq, A. Yang, A. Nagrani, P. Seo, A. Miech, J. Pont-Tuset, I. Laptev, J. Sivic, C. Schmid, CVPR 2023]

Vid2Seq – large-scale pretraining

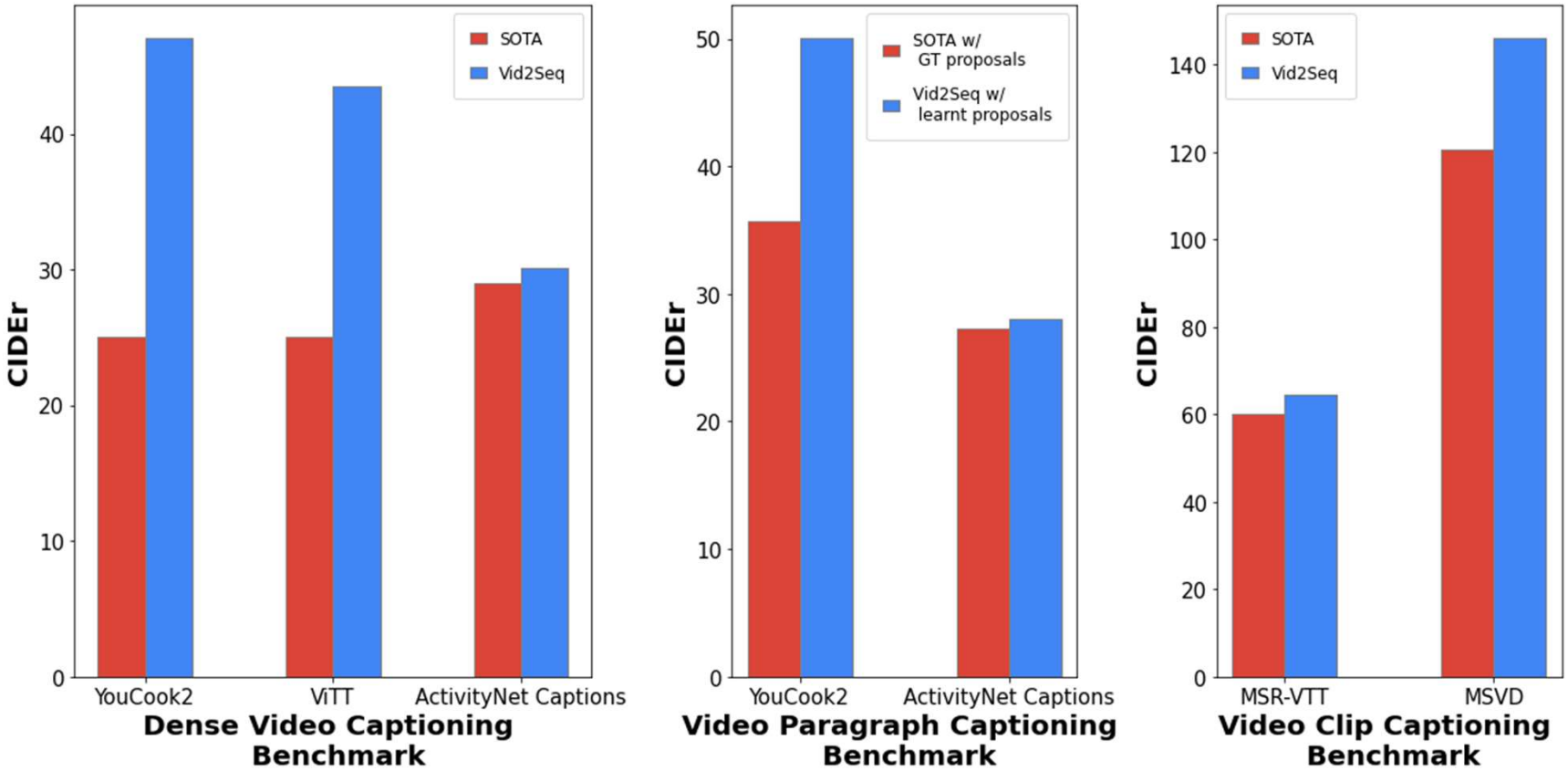
- Pretraining dataset is 15 million YouTube narrated videos from YT-Temporal-1B

- ASR sentence boundaries used as event boundaries

- Generative loss: given visual input predict speech
- Denoising loss: given visual input and masked speech, predict the masked tokens

[Vid2Seq, A. Yang, A. Nagrani, P. Seo, A. Miech, J. Pont-Tuset, I. Laptev, J. Sivic, C. Schmid, CVPR 2023]

Vid2Seq – SOTA results



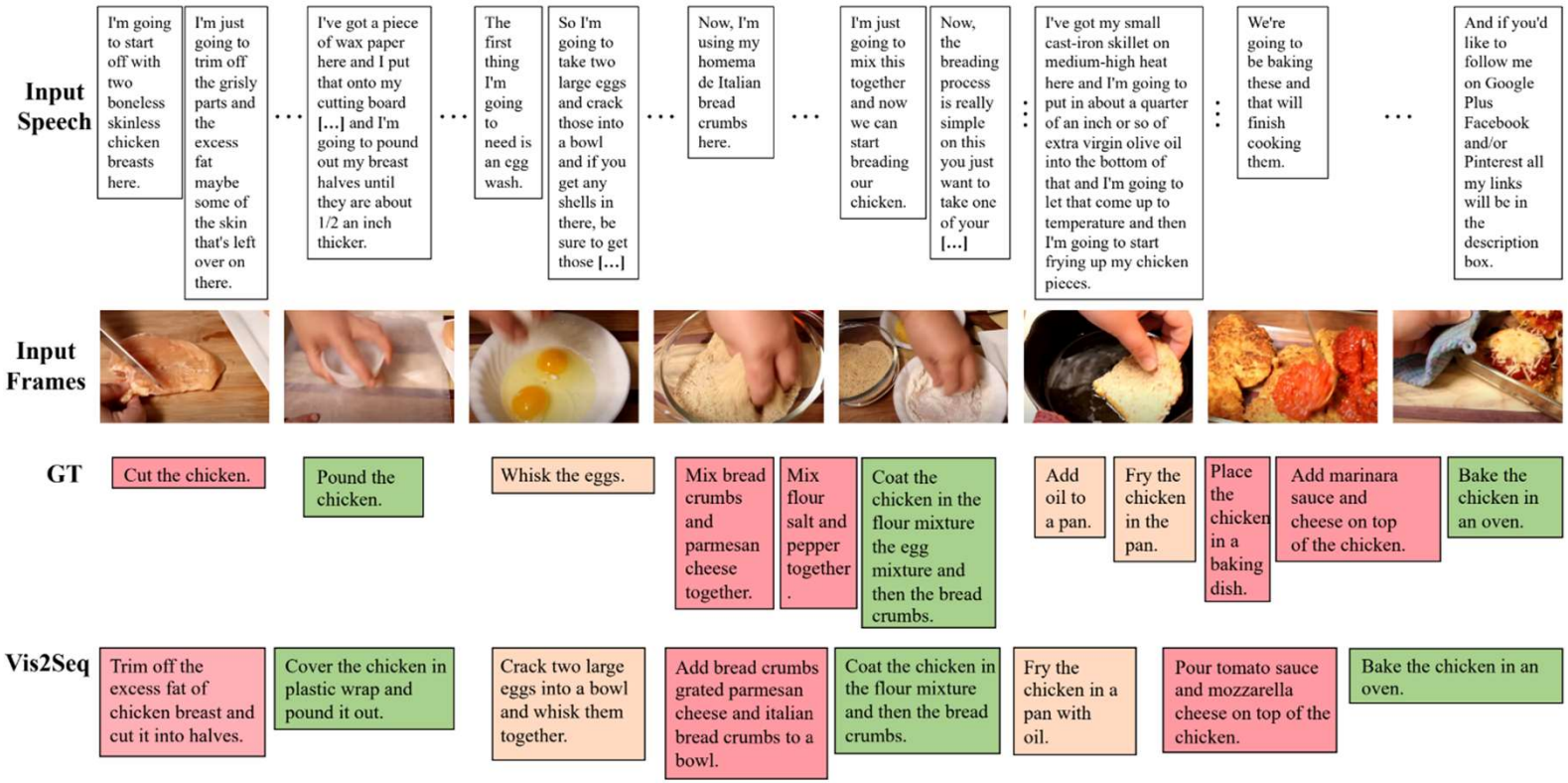
[Vid2Seq, A. Yang, A. Nagrani, P. Seo, A. Miech, J. Pont-Tuset, I. Laptev, J. Sivic, C. Schmid, CVPR 2023]

Ablation studies

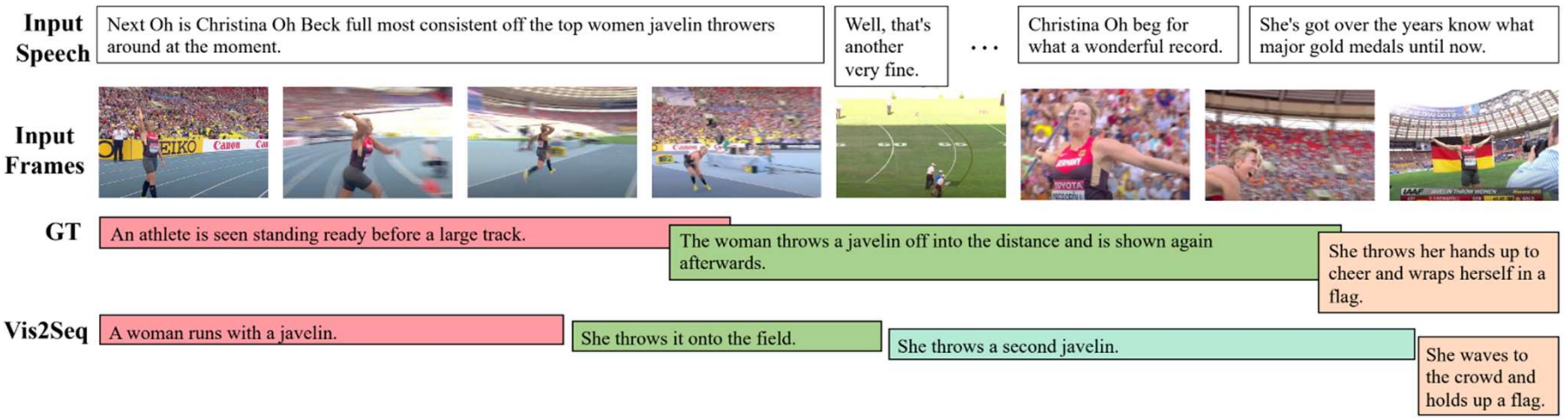
- Pretraining is important, datasize and quality matter
- Time tokens help when pretraining on untrimmed videos
- Visual and speech information is complementary
- Importance of losses: denoising loss is important if we use speech during pretraining

[Vid2Seq, A. Yang, A. Nagrani, P. Seo, A. Miech, J. Pont-Tuset, I. Laptev, J. Sivic, C. Schmid, CVPR 2023]

Qualitative results



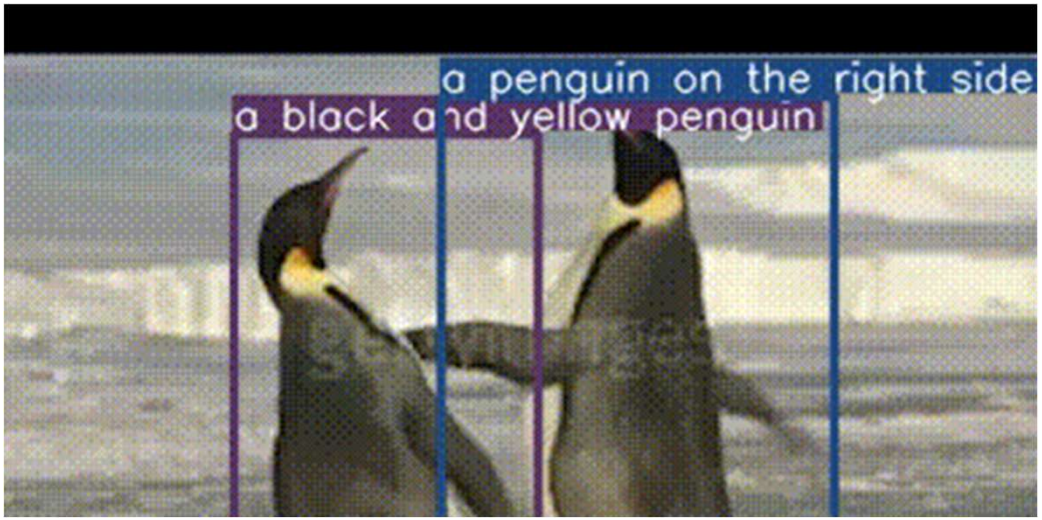
Qualitative results



Conclusion

- Excellent results for dense video captioning
- Metrics + variations on the detail of annotation
- Open issues
  - Small objects not detected & described by the caption
  - Rare events not detected
  - Missing use of internal and external knowledge

Outlook: Dense video object captioning

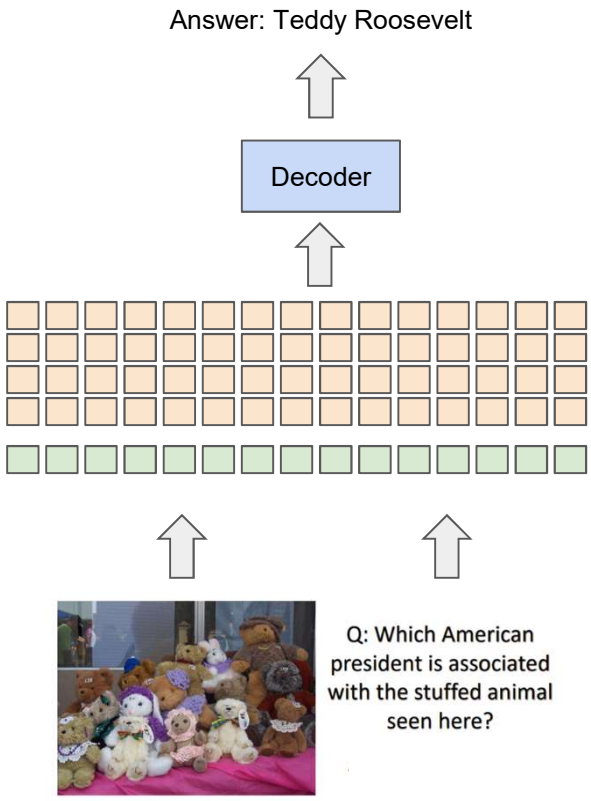


[Dense Video Object Captioning from Disjoint Supervision, X. Zhou, A. Arnab, C. Sun, C. Schmid, arXiv'23]

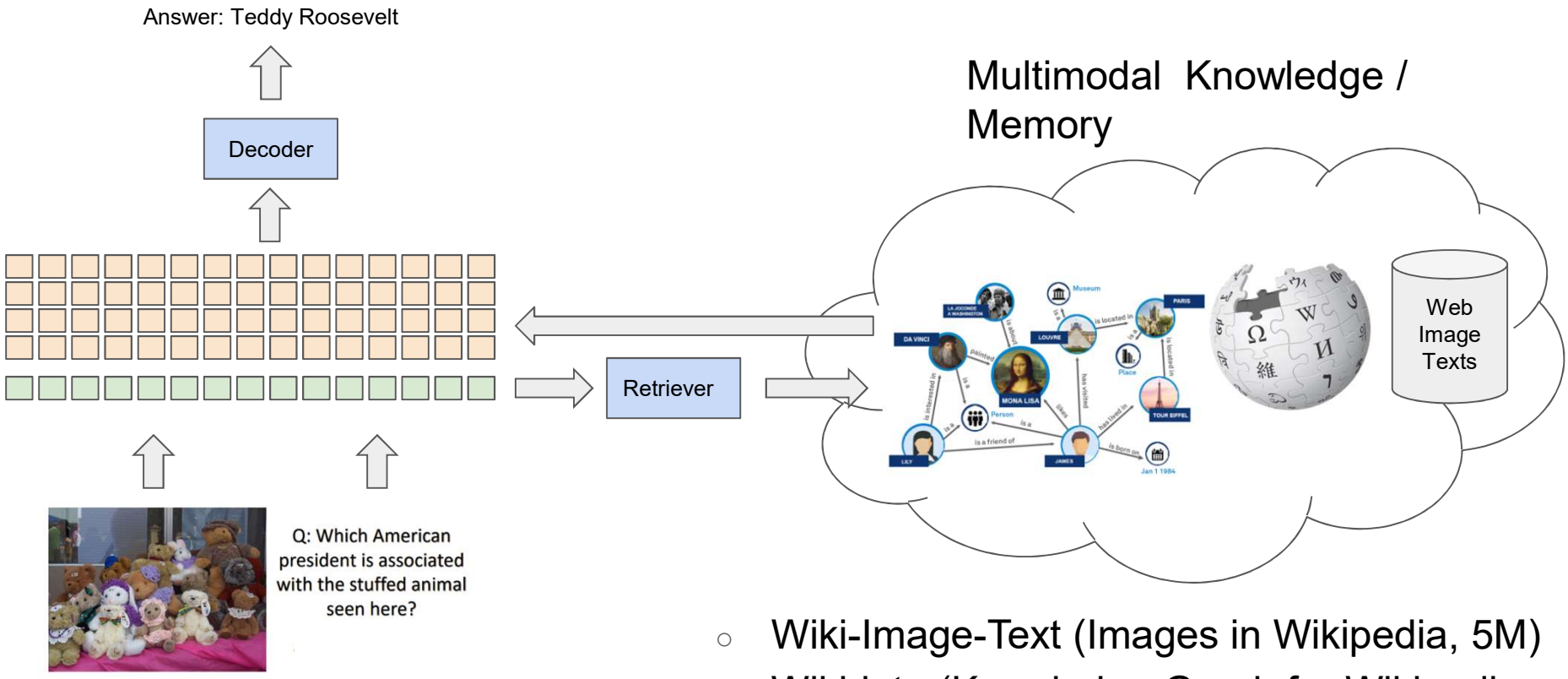
Overview

- VideoBERT
- Dense Video Captioning
- ***Retrieval Augmented Visual Question Answering***
- Multimodal transformer for navigation and manipulation

LLM with outside knowledge



LLM with outside knowledge



- Wiki-Image-Text (Images in Wikipedia, 5M)
- Wikidata (Knowledge Graph for Wikipedia entities, 12B triplets)

Why memory / knowledge?

- More accurate models: LLM are dedicated to high-level reasoning and memory to fine-grained and rare classes
- Disentangling knowledge from reasoning, use existing knowledge
- Retrieved memory / knowledge can be used to interpret model decisions
- Incremental learning w/o catastrophic forgetting: memory update without requiring to update the model

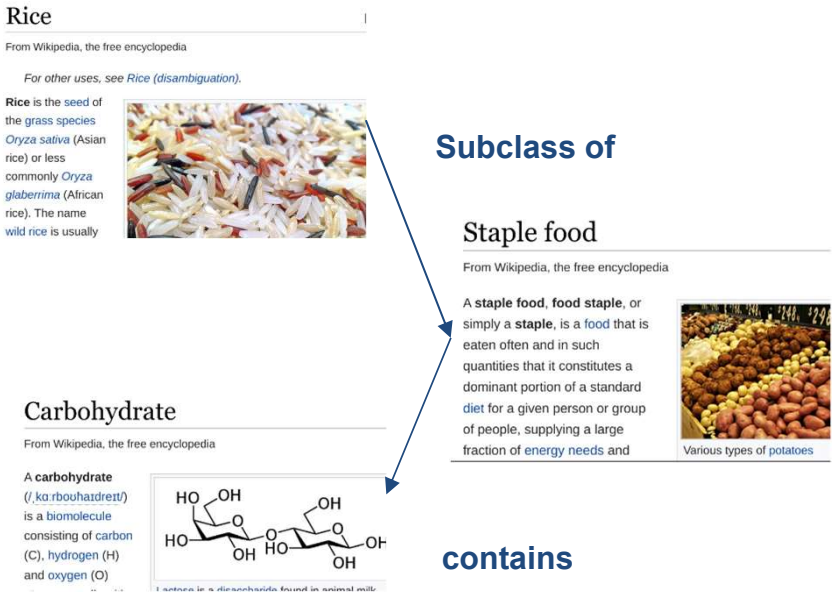
Why memory/ knowledge for VQA?

Answering the question requires additional information

Question : Which part of this meal has the most carbohydrates?

Answer: rice

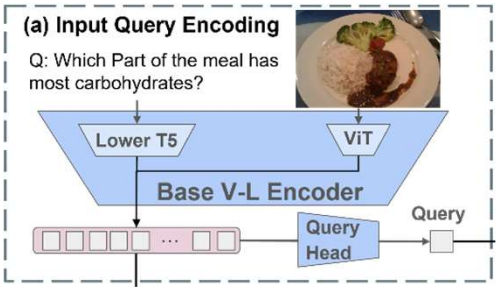
Example from OK-VQA



[REVEAL: Retrieval-Augmented Visual-Language Pre-Training with Multi-Source Multimodal Knowledge Memory, Z. Hu, A. Iscen, C. Sun, Z. Wang, K.-W. Chang, Y. Sun, C. Schmid, D. Ross, A. Fathi, CVPR'23]

Retrieval-augmented vision language model

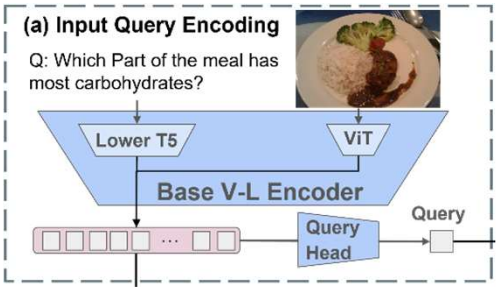
Model - Encoder

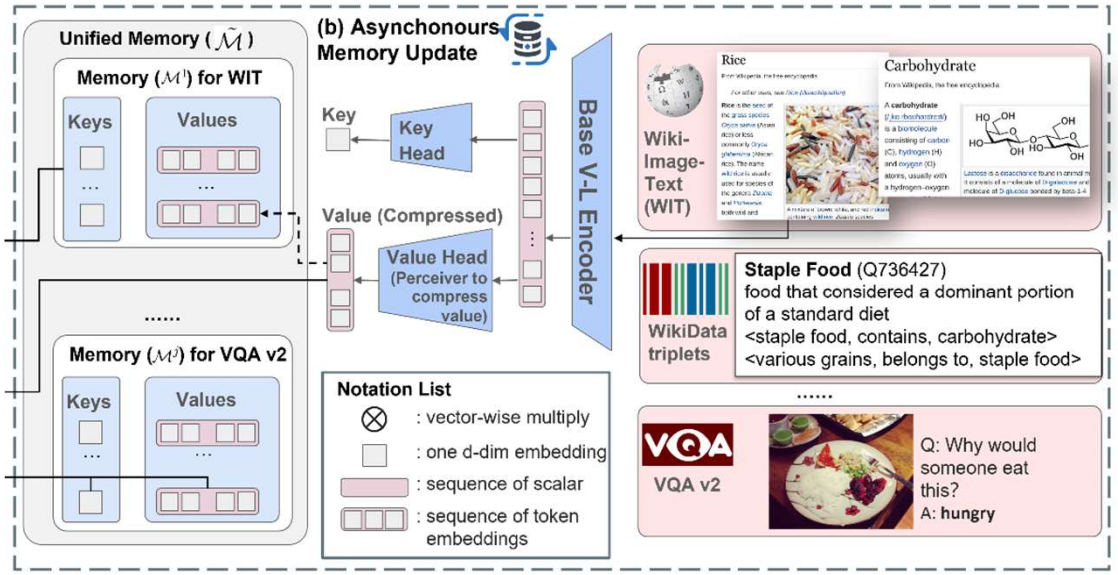


[REVEAL, Z. Hu, A. Iscen, C. Sun, Z. Wang, K.-W. Chang, Y. Sun, C. Schmid, D. Ross, A. Fathi, CVPR'23]

Retrieval-augmented vision language model

Model - Memory

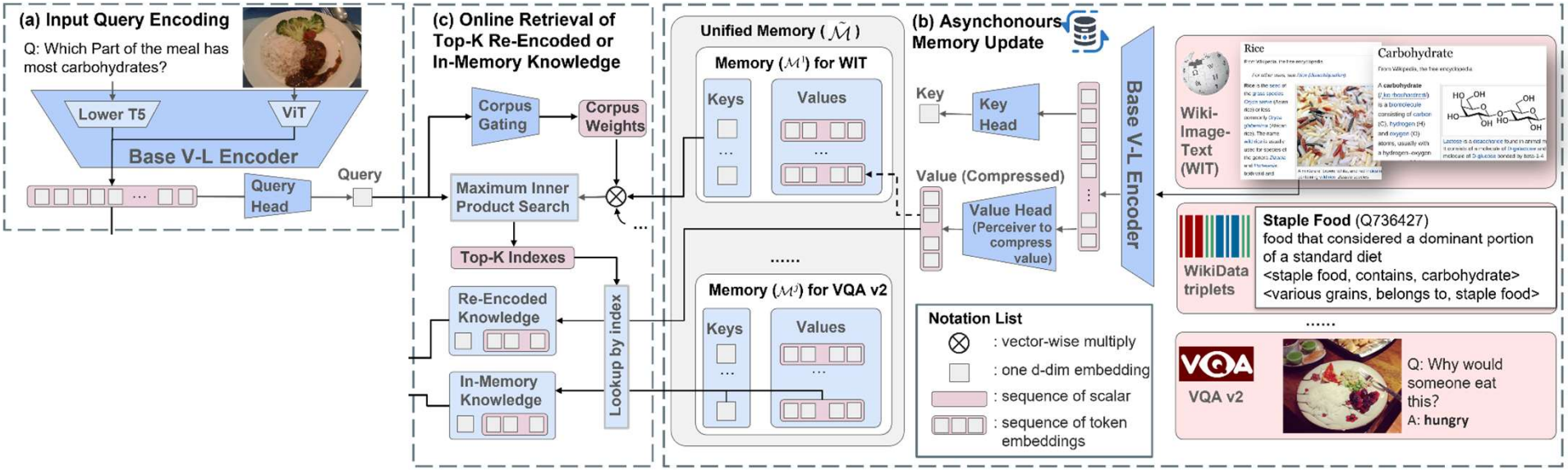




[REVEAL, Z. Hu, A. Iscen, C. Sun, Z. Wang, K.-W. Chang, Y. Sun, C. Schmid, D. Ross, A. Fathi, CVPR'23]

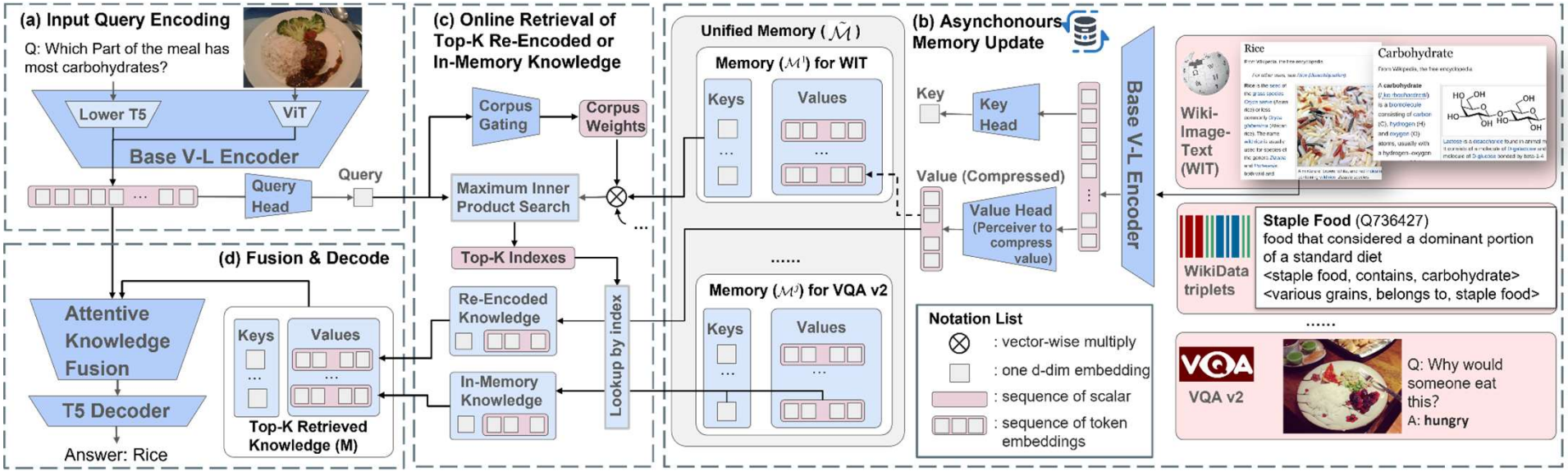
Retrieval-augmented vision language model

Model - Retriever



[REVEAL, Z. Hu, A. Iscen, C. Sun, Z. Wang, K.-W. Chang, Y. Sun, C. Schmid, D. Ross, A. Fathi, CVPR'23]

Retrieval-augmented vision language model

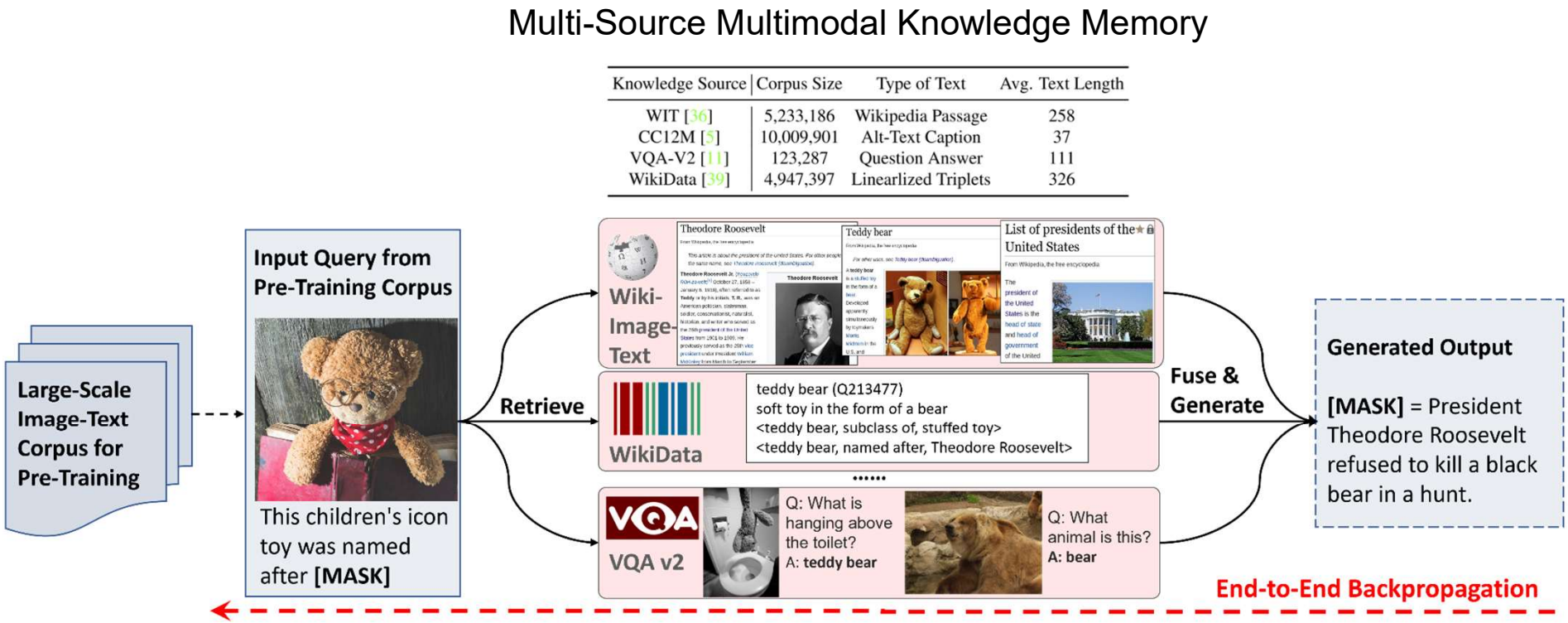


Model - Generator

[REVEAL, Z. Hu, A. Iscen, C. Sun, Z. Wang, K.-W. Chang, Y. Sun, C. Schmid, D. Ross, A. Fathi, CVPR'23]

Retrieval-augmented vision language model

Pretraining with image captioning



[REVEAL, Z. Hu, A. Iscen, C. Sun, Z. Wang, K.-W. Chang, Y. Sun, C. Schmid, D. Ross, A. Fathi, CVPR'23]

Results on OK-VQA

VQA Model Name	Knowledge Sources	Accuracy (%)	# params.
MUTAN+AN	Wikipedia + ConceptNet	27.8	-
ConceptBERT	Wikipedia	33.7	-
KRISP [27]	Wikipedia + ConceptNet	38.4	-
Visual Retriever-Reader	Google Search	39.2	-
MAVEx	Wikipedia+ConceptNet+Google Images	39.4	-
KAT-Explicit [12]	Wikidata	44.3	0.77B
PICa-Base [47]	Frozen GPT-3	43.3	(175B frozen)
PICa-Full [47]	Frozen GPT-3	48.0	(175B frozen)
KAT [12] (Single)	Wikidata + Frozen GPT-3	53.1	0.77B + (176B frozen)
KAT [12] (Ensemble)	Wikidata + Frozen GPT-3	54.4	2.31B + (176B frozen)
ReVIVE [23] (Single)	Wikidata + Frozen GPT-3	56.6	0.77B + (176.9B frozen)
ReVIVE [23] (Ensemble)	Wikidata+Frozen GPT-3	58.0	2.31B + (176.9B frozen)
REVEAL-Base	WIT + CC12M + Wikidata + VQA-2	55.2	0.4B
REVEAL-Large	WIT + CC12M + Wikidata + VQA-2	58.0	1.4B
REVEAL	WIT + CC12M + Wikidata + VQA-2	**59.1**	2.1B

Model Name	T5 Variant	Image Encoder	# params.	GFLOPs
REVEAL-Base	T5-Base	ViT-B/16	0.4B	120
REVEAL-Large	T5-Large	ViT-L/16	1.4B	528
REVEAL	T5-Large	ViT-g/14	2.1B	795

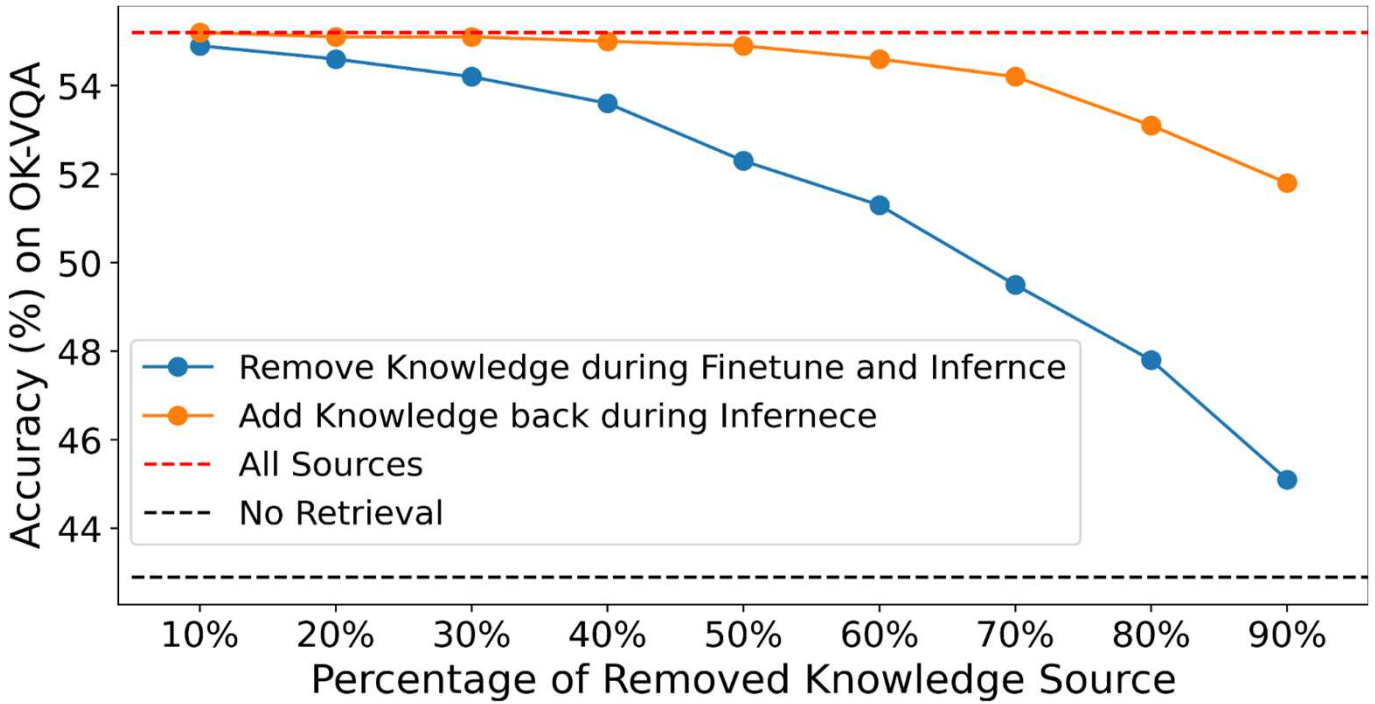
Table 2. Model configuration of different REVEAL variants.

[REVEAL, Z. Hu, A. Iscen, C. Sun, Z. Wang, K.-W. Chang, Y. Sun, C. Schmid, D. Ross, A. Fathi, CVPR'23]

Example results

[REVEAL, Z. Hu, A. Iscen, C. Sun, Z. Wang, K.-W. Chang, Y. Sun, C. Schmid, D. Ross, A. Fathi, CVPR'23]

How useful is knowledge memory?



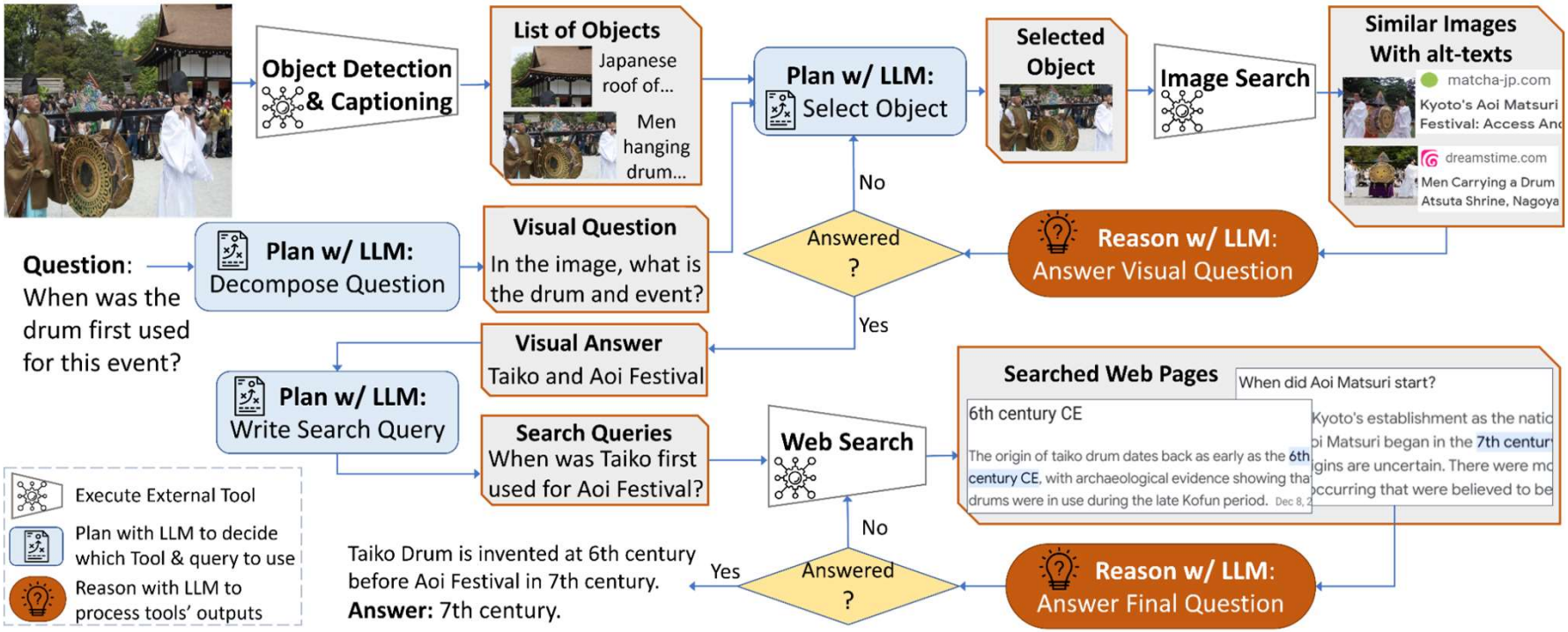
Blue curve: x% removed during fine-tuning and inference
Orange curve: x% removed during fine-tuning, but added during inference; this simulates on-the-fly knowledge update

[REVEAL, Z. Hu, A. Iscen, C. Sun, Z. Wang, K.-W. Chang, Y. Sun, C. Schmid, D. Ross, A. Fathi, CVPR'23]

Conclusion

- Excellent performance with large-scale models
- Importance of large-scale, but also high quality training data
- Retrieval augmentations improves and complements large-scale models

Outlook: Large language model for reasoning



Example of generated workflow:

- LLM-based planner the dynamically selects the external tool
- LLM-based reasoner to process tool output

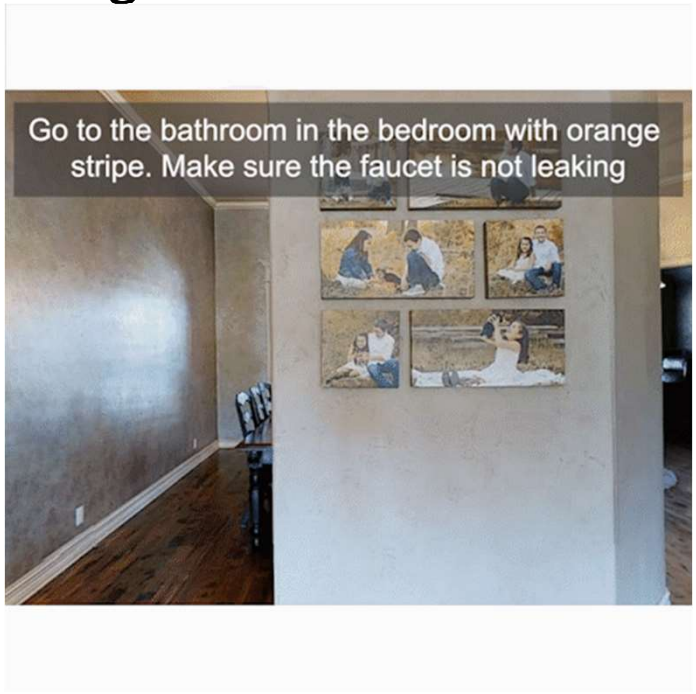
[AVIS: Autonomous Visual Information Seeking with Large Language Model Agent, Z. Hu, A. Iscen, C. Sun, K.-W. Chang, Y. Sun, D. Ross, C. Schmid, A. Fathi, Neurips'23]

Overview

- VideoBERT
- Dense Video Captioning
- Retrieval Augmented Visual Question Answering
- ***Multimodal transformer for navigation & manipulation***

History Aware Multimodal Transformer (HAMT) for Vision-and-Language Navigation

- Train autonomous agent to follow natural language instructions to navigate in in-door environments

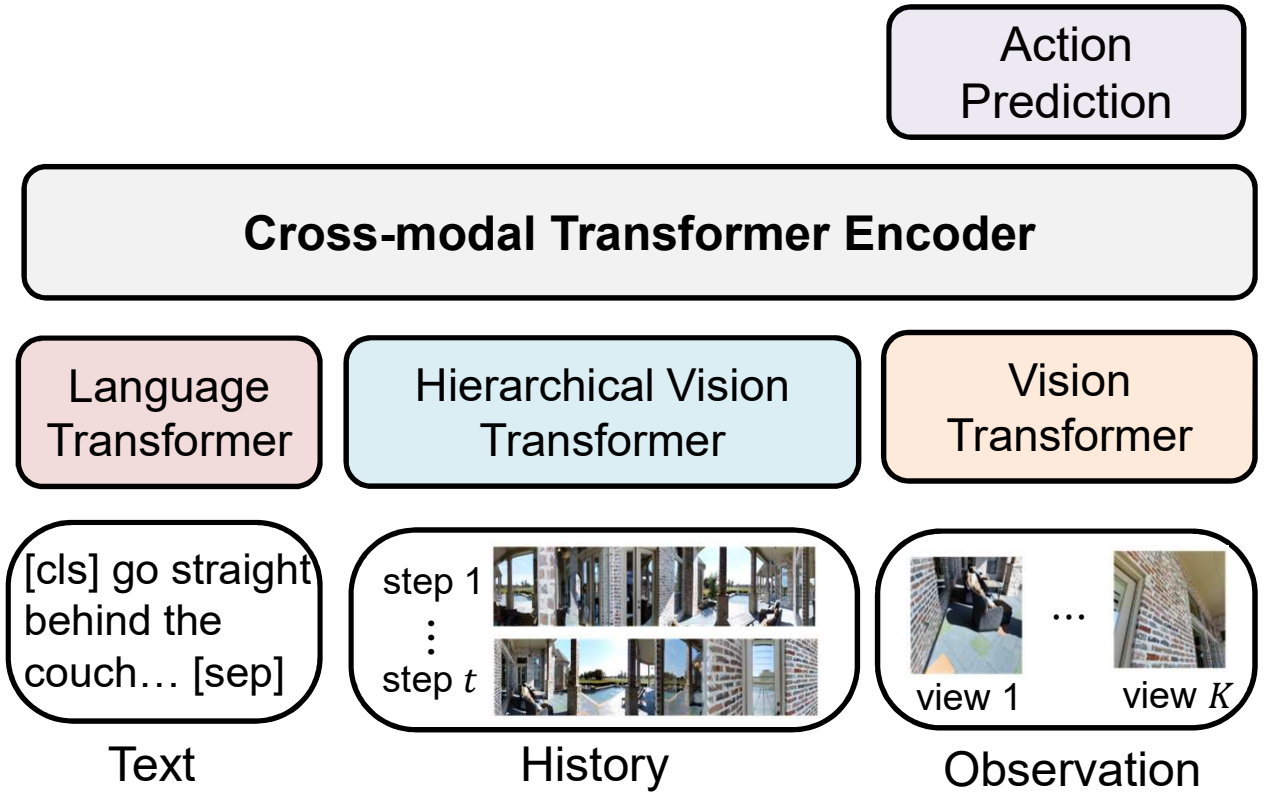


- Design of a fully transformer based model

[HAMT, S. Chen, PL. Guhur, C. Schmid, I. Laptev, Neurips'21]

History Aware Multimodal transformer (HAMT)

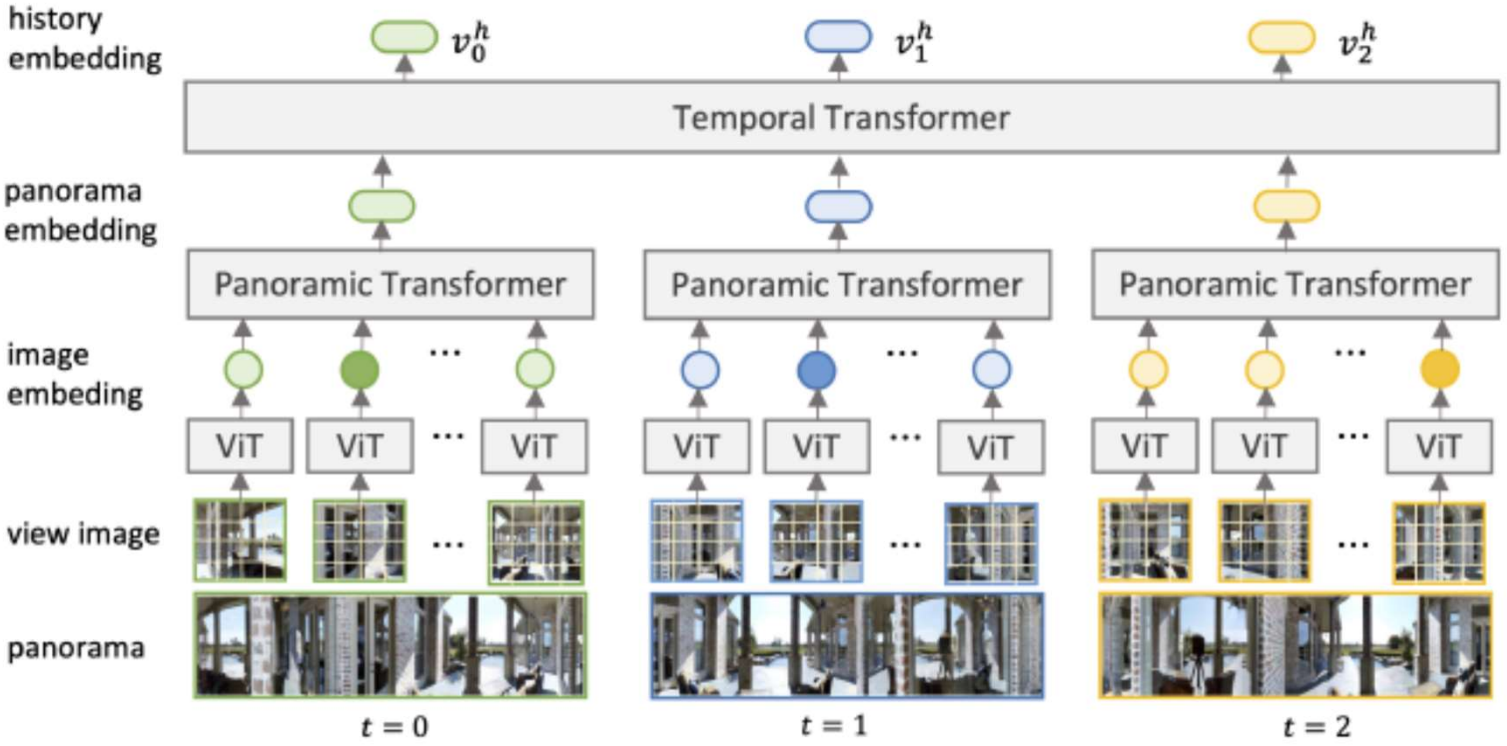
- Long-horizon history modeling for learning temporal dependency of observations and actions



[HAMT, S. Chen, PL. Guhur, C. Schmid, I. Laptev, Neurips'21]

Hierarchical history encoding

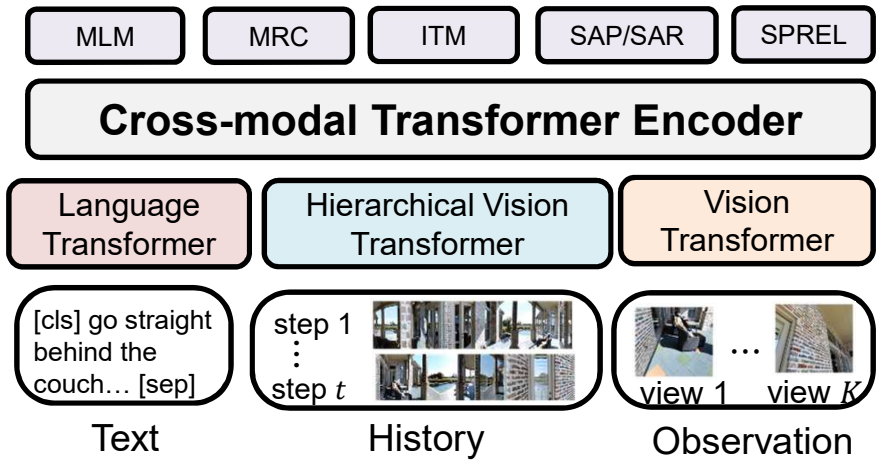
- ViT for single view image encoding
- Panoramic transformer for view encoding
- Temporal transformer for sequence encoding



[HAMT, S. Chen, PL. Guhur, C. Schmid, I. Laptev, Neurips'21]

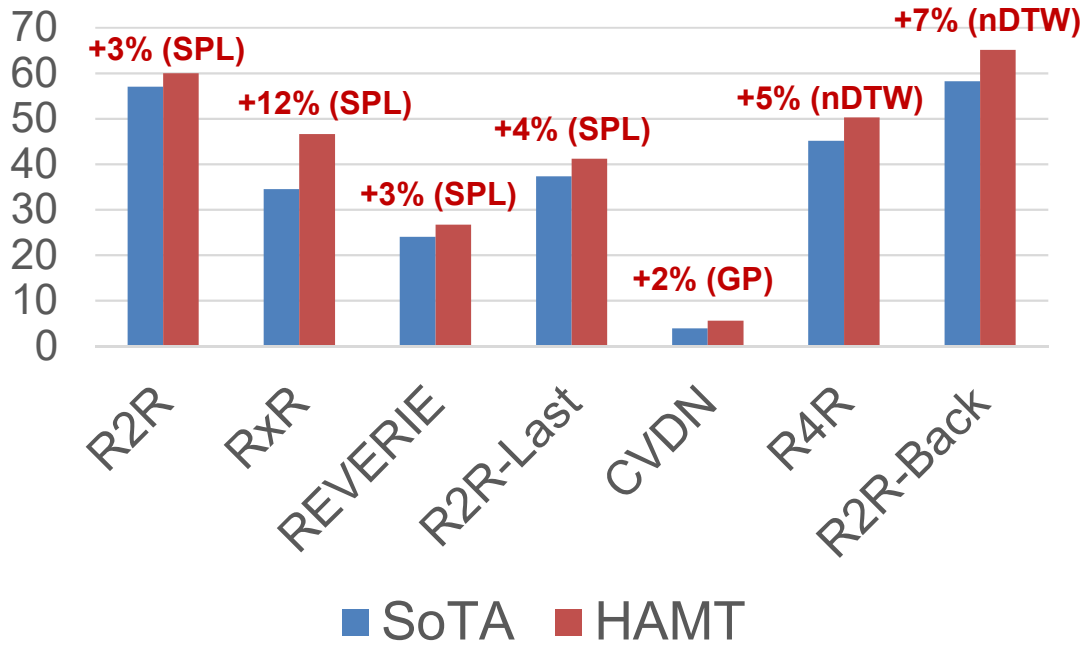
End-to-end training with several losses

- Common vision-and-language proxy tasks
  - Masked Language Modelling
  - Masked Region Modelling
  - Instruction Trajectory Matching
- New proxy tasks for VLN
  - Single-step Action Regression
  - Spatial Relationship Prediction

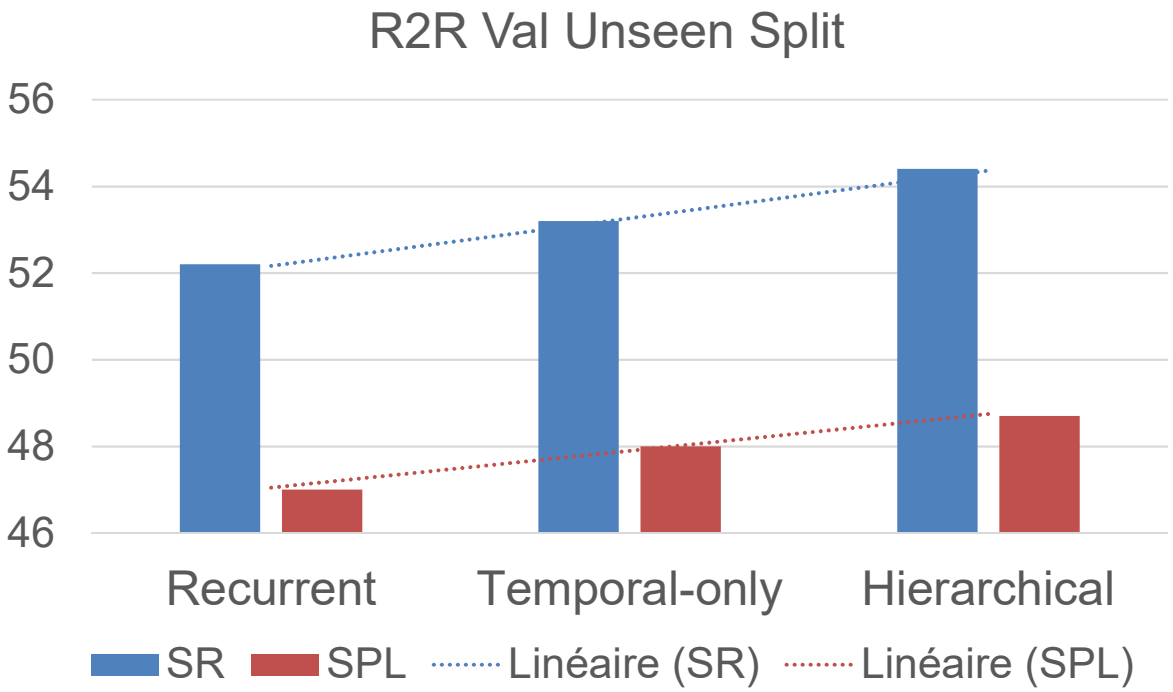


Experiments: Comparison with SOTA

- HAMT outperforms state of the art

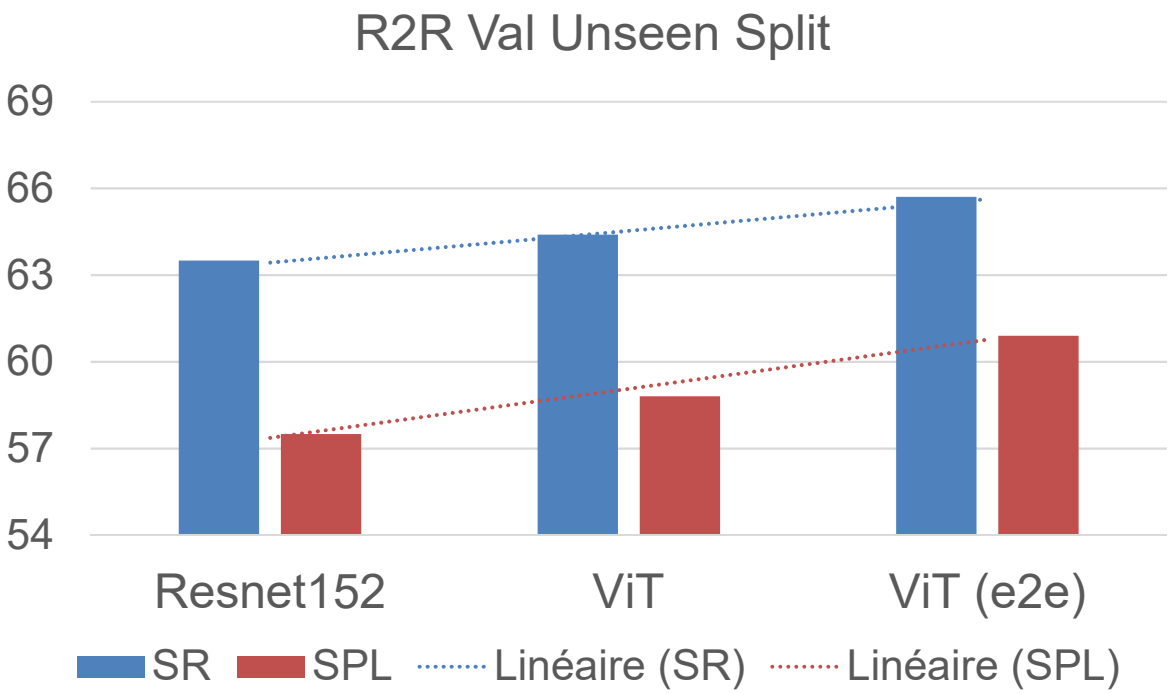


Ablation: transformer / hierarchical model



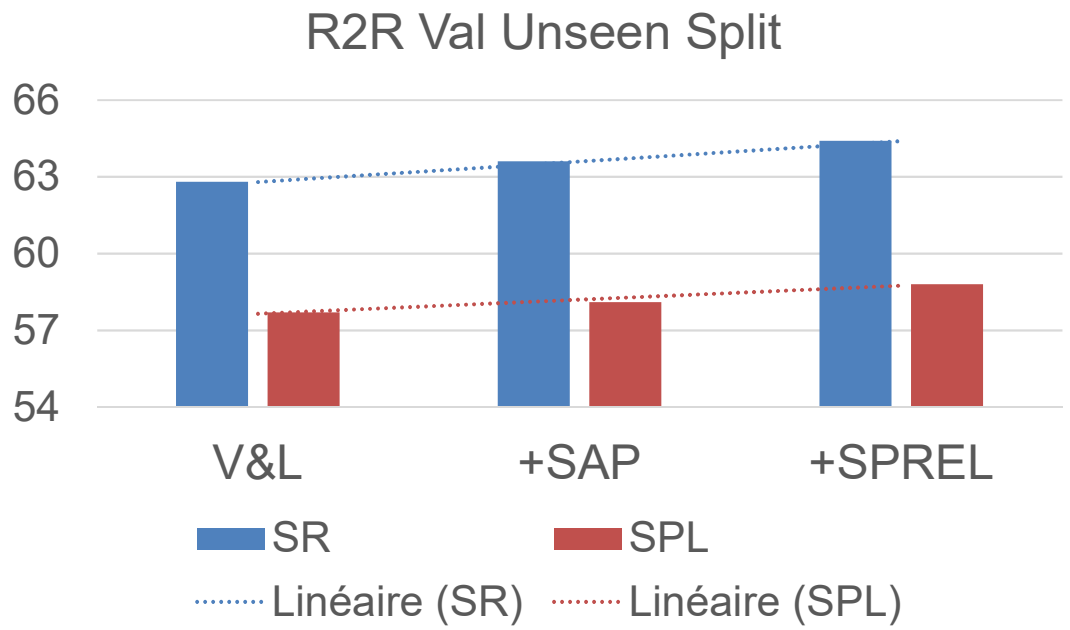
- Recurrent: a fixed-size vector to encode the whole history
- Temporal-only: transformer / one view per panorama to improve efficiency
- **Hierarchical: transformer with hierarchically encoded panorama**

Ablation: visual representation



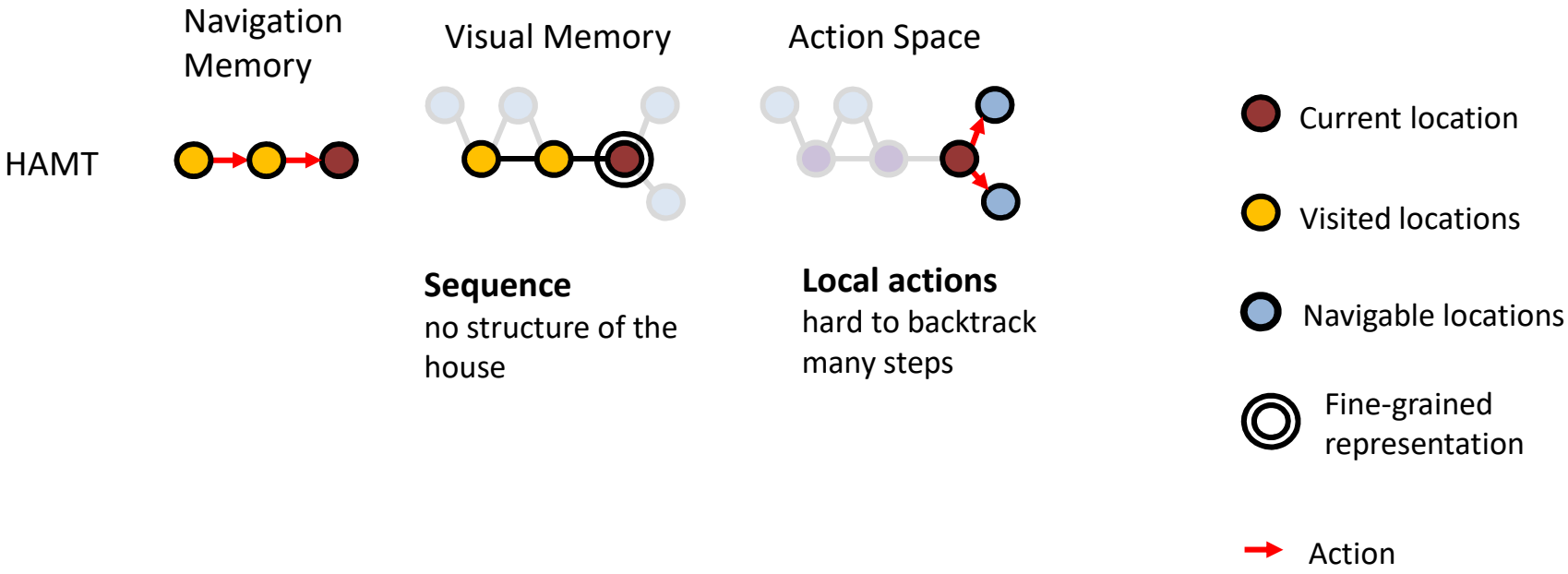
- Resnet152 pretrained on ImageNet
- ViT pretrained on ImageNet
- **ViT e2e optimized on VLN task**

Ablation: training with proxy tasks



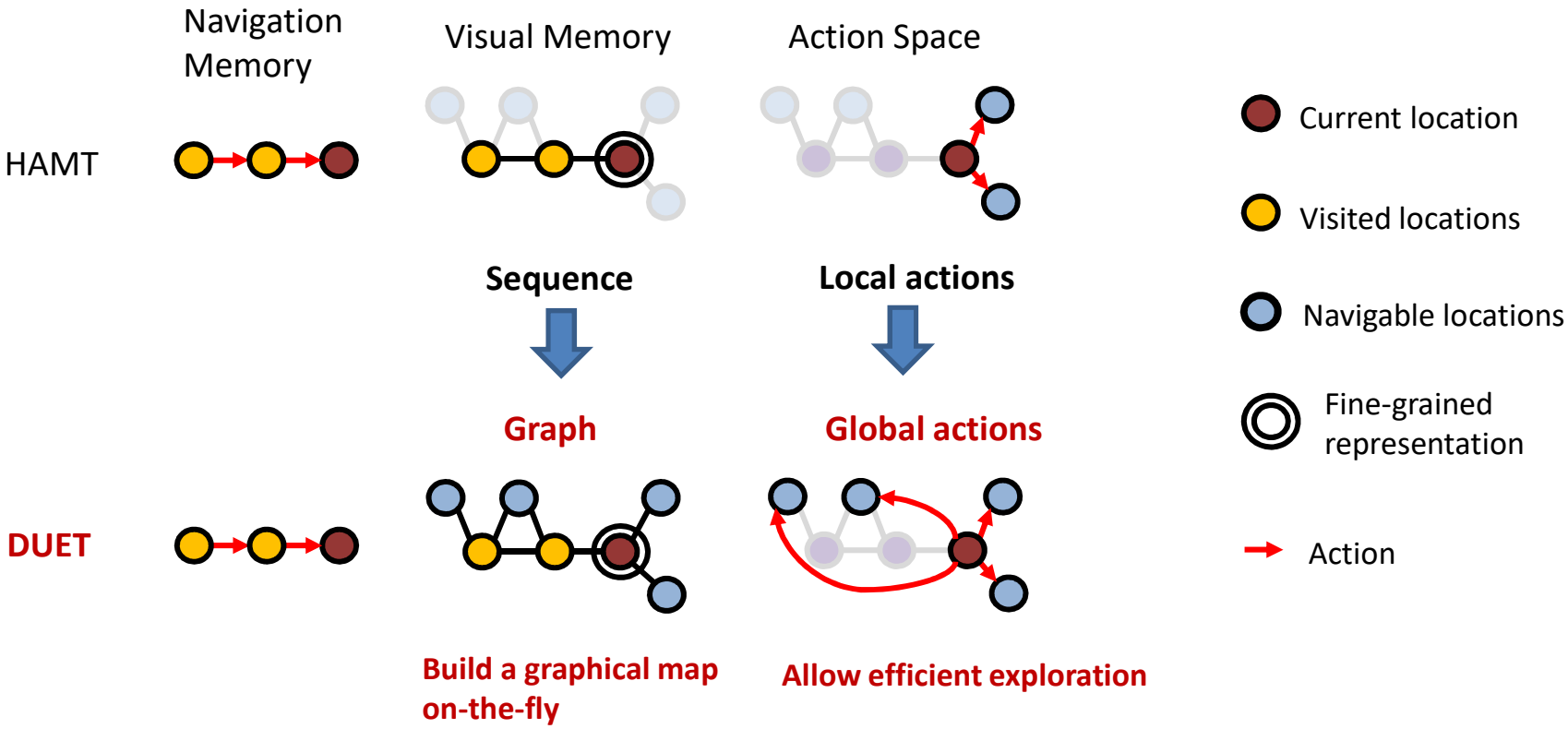
- Common V&L proxy tasks
- + Single-step action prediction (SAP)
- + Spatial relation prediction (SPREL)

Limitations of HAMT



- Sequential representation, doesn't learn a structure of the house
- Only local actions, hard to backtrack

Improving HAMT with Structured Memory



[DUET, S. Chen, PL. Guhur, M. Tapaswi, C. Schmid, I. Laptev, CVPR'22]

DUET: Experimental Results

- REVERIE dataset

	SR	SPL	RGS	RGS PL
HAMT	30.40	26.67	14.88	13.08
DUET	**52.51**	**36.06**	**31.88**	**22.06**

- SOON dataset

Split	Methods	TL	OSR↑	SR↑	SPL↑	RGSPL↑
Val Unseen	GBE [8]	28.96	28.54	19.52	13.34	1.16
	DUET (Ours)	36.20	**50.91**	**36.28**	**22.58**	**3.75**
Test Unseen	GBE [8]	27.88	21.45	12.90	9.23	0.45
	DUET (Ours)	41.83	**43.00**	**33.44**	**21.42**	**4.17**

- **Winner of VLN Challenges** hosted in Human Interaction for Robotics Navigation Workshop at ICCV 2021


1ST PLACE IN THE
REVERIE CHALLENGE 2021

Shizhe Chen[1], Pierre-Louis Guhur[1], Makarand Tapaswi[2]
Cordelia Schmid[1] and Ivan Laptev[1]

[1]Inria, École normale supérieure, CNRS, PSL Research University
[2]IIIT Hyderabad

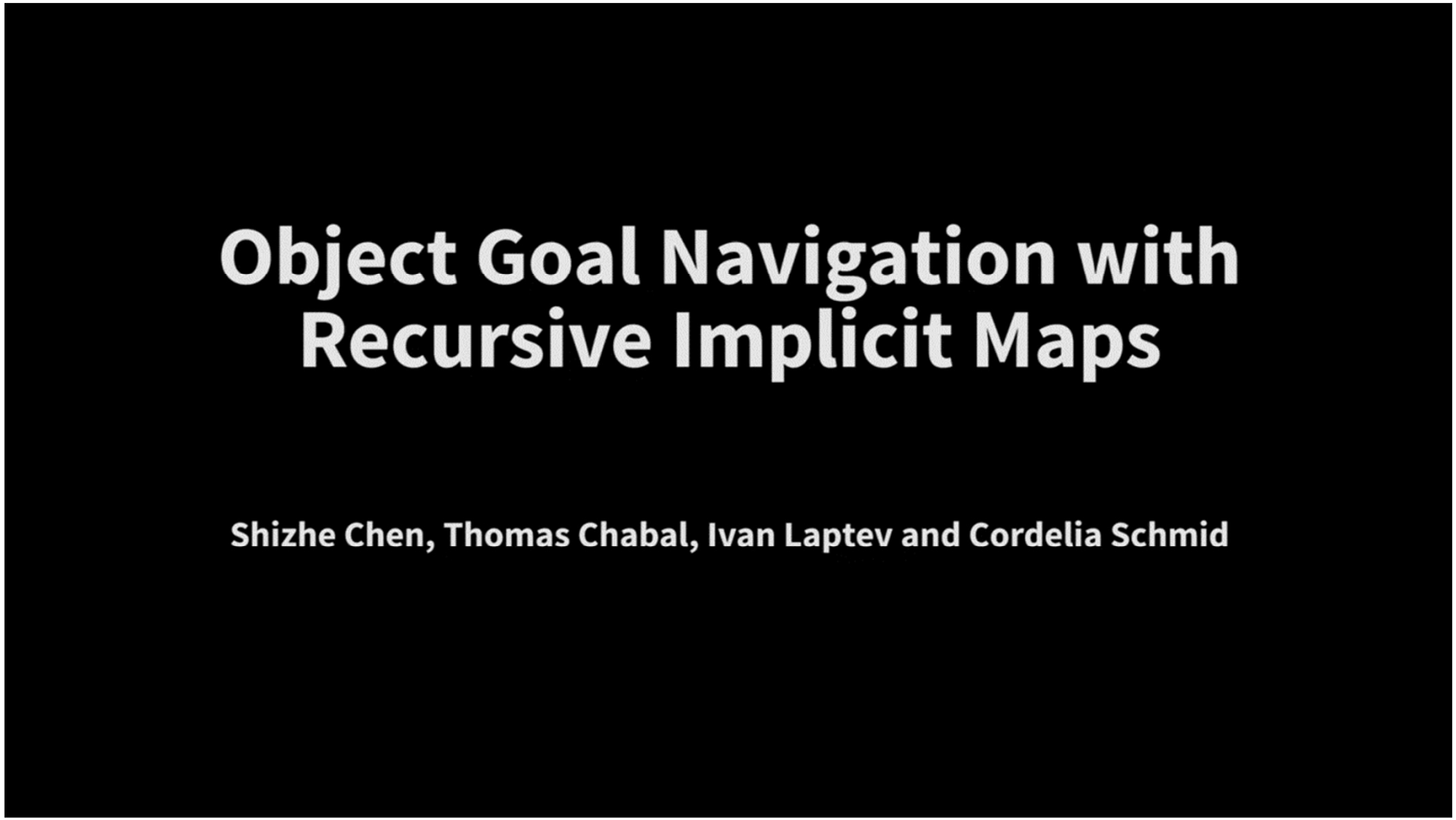
presented at the

Human Interaction for Robotic Navigation Workshop
at the IEEE/CVF International Conference on Computer Vision (ICCV) 2021

Oct. 16
2021

SIGNED, Dr. Qi Wu
On behalf of the 2020 REVERIE Challenge Organizers

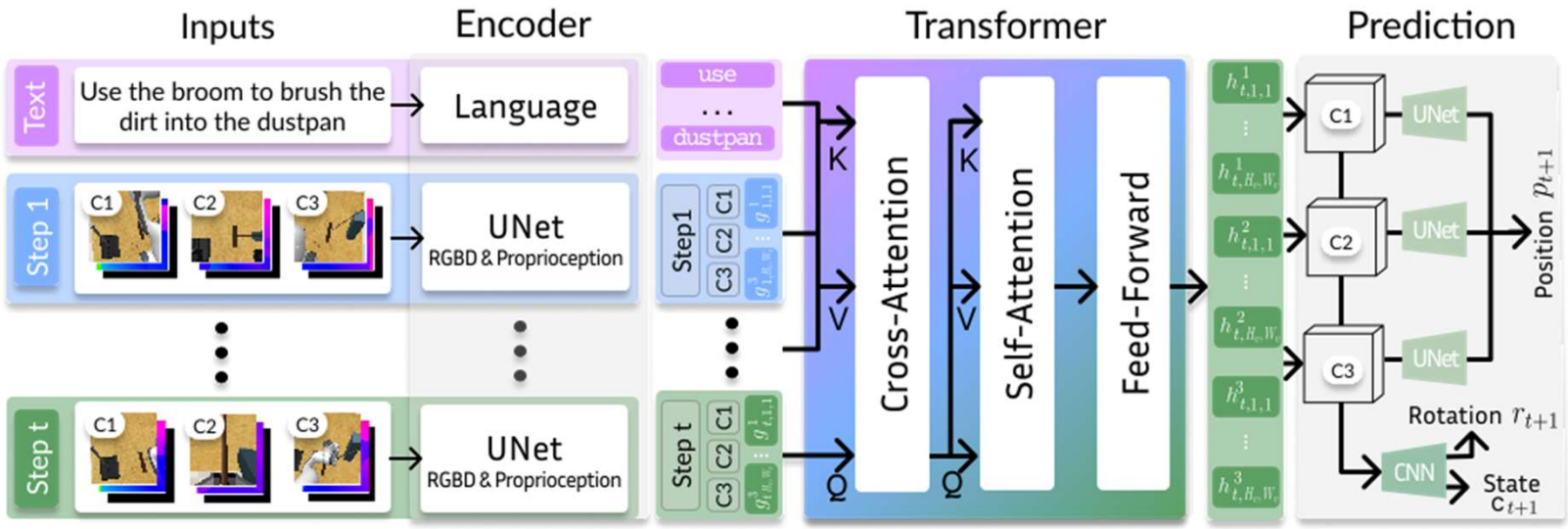
Yuankai Qi Fengda Zhu Qi Wu



[Object Goal Navigation with Recursive Implicit Maps, S. Chen, T. Chabel, I. Laptev, C. Schmid, IROS 2023]

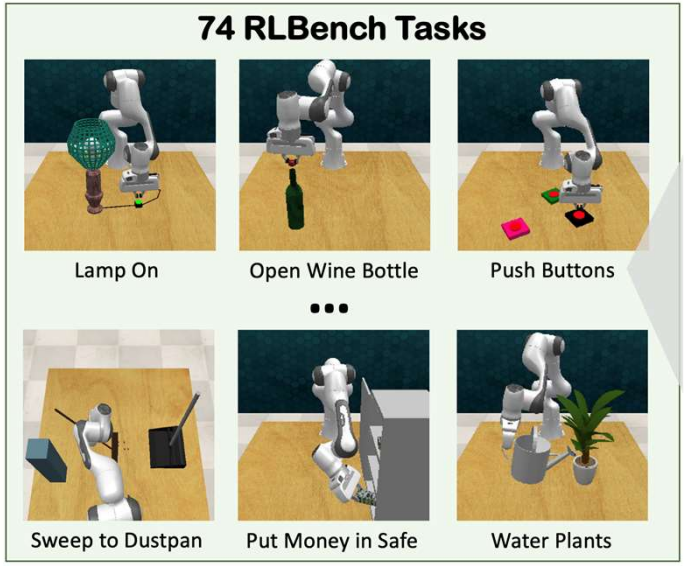
Hiveformer for robot manipulation

- Hiveformer: History-aware Instruction-conditioned multi-view transformer
  - use all the previous multi-view RGB-D images and actions in self-attention
  - learn cross-modal alignment between text, image and action in cross-attention



[Hiveformer, PL. Guhur, S. Chen, I. Laptev, C. Schmid, I. Laptev, CORL'22]

Evaluation: RLBench



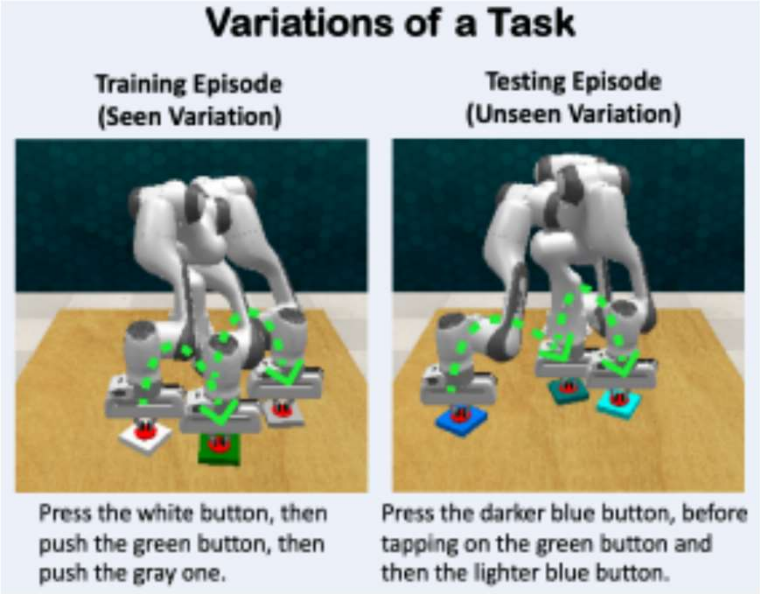
- 100 hand-designed tasks
- Multi-view RGB-D images, Franka 7 DoF arm
- Text description for each task

⇨ Select 74 tasks we could simulate

⇨ Evaluate in single and multi-task settings

[RLBench, S. James, Z. Ma, D. Arrojo, A. Davison, IEEE RAL'20]

Evaluation: RLBench task variations



Unseen sequence of colors during training

Evaluate on

unseen task variations

Results: 74 tasks • Single-task setting

- Manually group 74 RLBench tasks into 9 groups

	Planning	Tools	Long Term	Rot. Invar.	Motion Planning	Screw	Multi Modal	Precision	Visual Occlusion	Avg
Num. of tasks	9	11	4	7	9	4	5	11	14	74
Auto-λ [14]	58.9	20.0	2.3	73.1	66.7	48.2	47.6	34.6	40.6	44.0
Ours (w/o hist)	78.9	46.7	10.0	84.6	73.3	72.6	60.0	63.8	57.9	60.9
Ours (one view)	57.7	23.2	12.3	57.8	63.2	35.6	40.7	33.7	37.1	40.1
Ours	81.6	53.0	16.9	84.2	72.7	80.9	67.1	64.7	60.2	65.4

+21.4%

- HiveFormer generalizes well to many tasks: +21.4% over [14]
- History matters especially **Planning**, **Tools** and **Long-Terms** tasks
- Multi-view matters especially for **Screw**, **Precision** and **Visual Occlusion** tasks

[Hiveformer, PL. Guhur, S. Chen, I. Laptev, C. Schmid, I. Laptev, CORL'22]

Results: 74 tasks • Single-task setting

- Manually group 74 RLBench tasks into 9 groups

	Planning	Tools	Long Term	Rot. Invar.	Motion Planning	Screw	Multi Modal	Precision	Visual Occlusion	Avg
Num. of tasks	9	11	4	7	9	4	5	11	14	74
Auto-λ [14]	58.9	20.0	2.3	73.1	66.7	48.2	47.6	34.6	40.6	44.0
Ours (w/o hist)	78.9	46.7	10.0	84.6	73.3	72.6	60.0	63.8	57.9	60.9
Ours (one view)	57.7	23.2	12.3	77.8	63.2	35.6	40.7	33.7	37.1	40.1
Ours	81.6	53.0	16.9	84.2	72.7	80.9	67.1	64.7	60.2	65.4

- HiveFormer generalizes well to many tasks: +21.4% over [14]
- History matters especially **Planning**, **Tools** and **Long-Terms** tasks
- Multi-view matters especially for **Screw**, **Precision** and **Visual Occlusion** tasks

[Hiveformer, PL. Guhur, S. Chen, I. Laptev, C. Schmid, I. Laptev, CORL'22]

Results: 74 tasks • Single-task setting

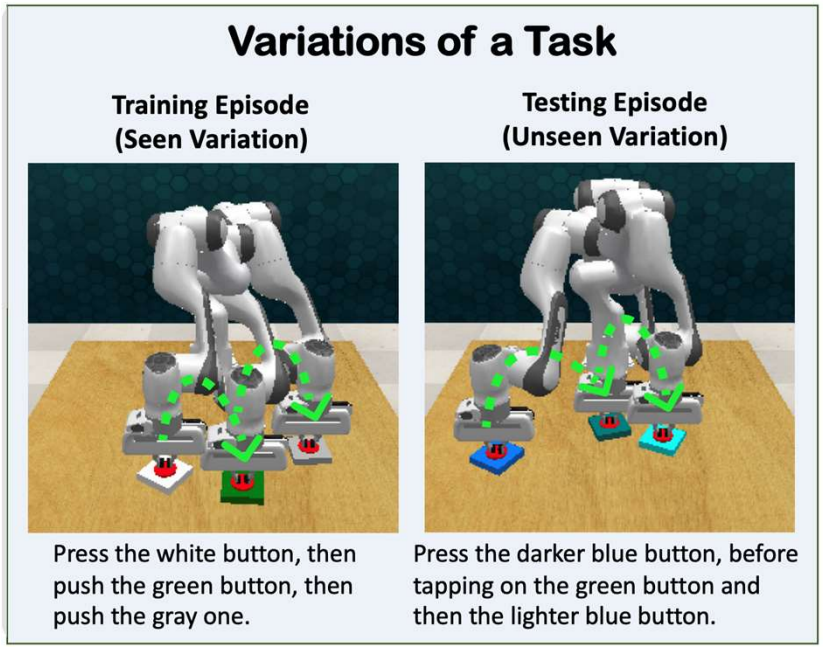
- Manually group 74 RLBench tasks into 9 groups

	Planning	Tools	Long Term	Rot. Invar.	Motion Planning	Screw	Multi Modal	Precision	Visual Occlusion	Avg
Num. of tasks	9	11	4	7	9	4	5	11	14	74
Auto-λ [14]	58.9	20.0	2.3	73.1	66.7	48.2	47.6	34.6	40.6	44.0
Ours (w/o hist)	78.9	46.7	10.0	84.6	73.3	72.6	60.0	63.8	57.9	60.9
Ours (one view)	57.7	23.2	12.3	57.8	63.2	35.6	40.7	33.7	37.1	40.1
Ours	81.6	53.0	16.9	84.2	72.7	80.9	67.1	64.7	60.2	65.4

- HiveFormer generalizes well to many tasks: +21.4% over [14]
- History matters especially **Planning**, **Tools** and **Long-Terms** tasks
- Multi-view matters especially for **Screw**, **Precision** and **Visual Occlusion** tasks

[Hiveformer, PL. Guhur, S. Chen, I. Laptev, C. Schmid, I. Laptev, CORL'22]

Results: Task variations

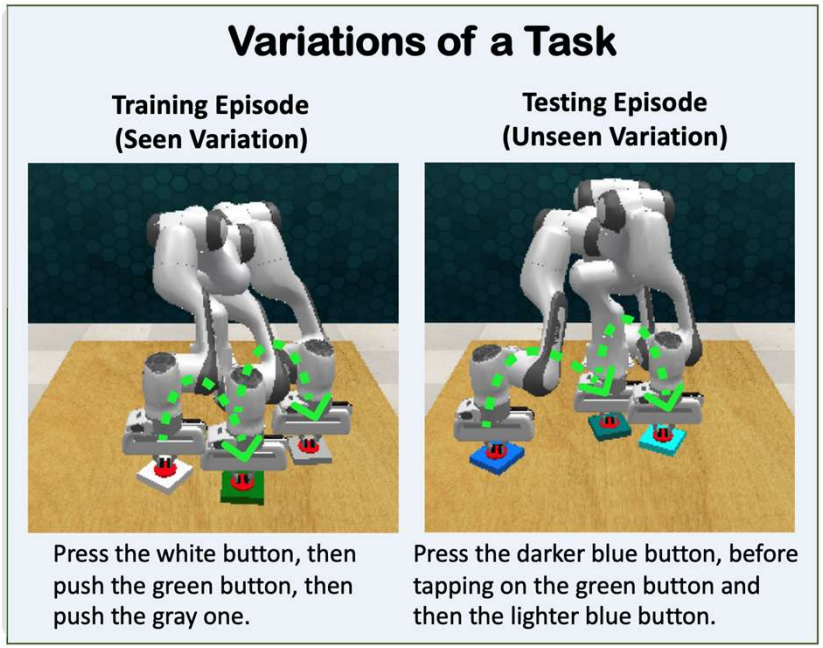


# Demos Per Variation	Instr.	Push Buttons Seen Synt.	Push Buttons Unseen Synt.	Push Buttons Unseen Real	Tower Seen Synt.	Tower Unseen Synt.	Tower Unseen Real
10	Seq.	96.4	71.1	65.7	71.6	49.8	19.4
50	Seq.	99.4	83.1	70.9	74.3	52.1	20.6
100	Seq.	100	86.3	74.2	77.4	56.2	24.1

- Generalization to unseen variations
- Generalization to natural language extractions

[Hiveformer, PL. Guhur, S. Chen, I. Laptev, C. Schmid, I. Laptev, CORL'22]

Results: Task variations



# Demos Per Variation	Instr.	Push Buttons Seen Synt.	Push Buttons Unseen Synt.	Push Buttons Unseen Real	Tower Seen Synt.	Tower Unseen Synt.	Tower Unseen Real
10	Seq.	96.4	71.1	65.7	71.6	49.8	19.4
50	Seq.	99.4	83.1	70.9	74.3	52.1	20.6
100	Seq.	100	86.3	74.2	77.4	56.2	24.1

- Generalization to unseen variations
- Generalization to natural language extractions

[Hiveformer, PL. Guhur, S. Chen, I. Laptev, C. Schmid, I. Laptev, CORL'22]

Hiveformer – robot manipulation

- SOTA on RLBench in single task and multi-task setting

- Real-robot experiments:
  - training on a small number of real-world demonstration
  - pre-training on RLBench results in a significant gain

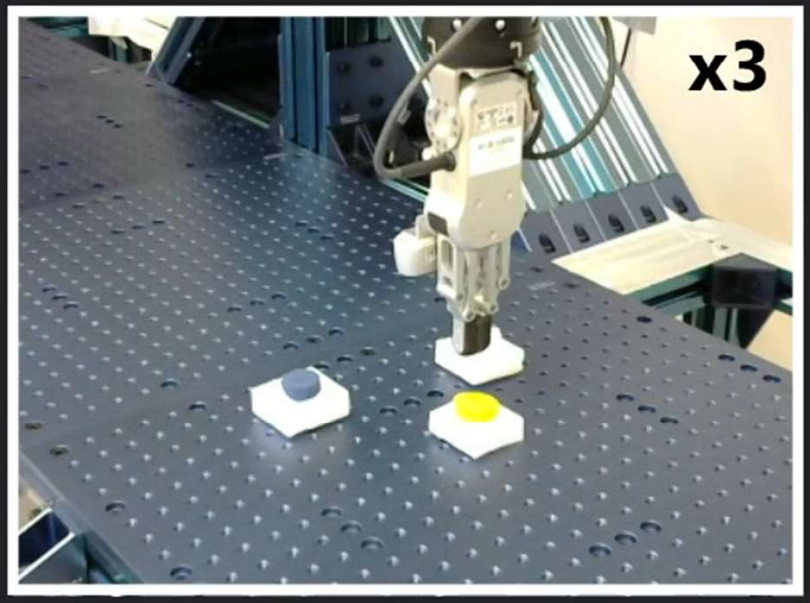
Pretrain	Seen Vars	Unseen Vars
-	86.7	13.3
✓	92.2	85.7

Success rate of push button task on real robot

[Hiveformer, PL. Guhur, S. Chen, I. Laptev, C. Schmid, I. Laptev, CORL'22]

Push Buttons - Success Case

Seen Variation - Synthetic Instruction



Push the yellow button, then press the white button, then push the blue button

Correct execution for varied instructions:

- Stack the yellow cup on top of the pink cup
- Put the yellow cup on top of the pink one
- Place the yellow cup onto the pink one

Correct execution for varied instructions:

- Put a strawberry, then a lemon, then a strawberry in the box
- Take the strawberry and put it in the box, then take a lemon and put it in the box, then put another strawberry in the box

Conclusion

- Transformers can be used to predict actions
- Standard vision-language losses improve performance, can leverage existing vision-language models
- Pretraining data useful
- Excellent results on benchmarks, initial results on a real robot promising

THANK YOU