MOHAMED BIN ZAYED
UNIVERSITY OF
ARTIFICIAL INTELLIGENCE

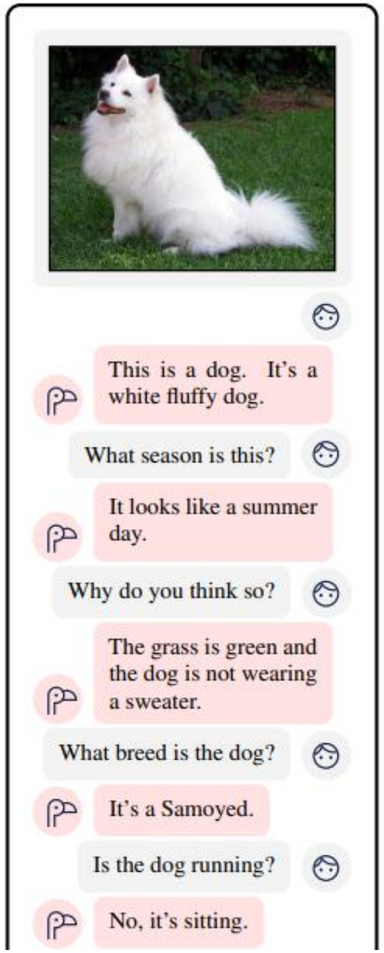
International Summer School on Machine Vision
Padova, Italy - September 4-8, 2023

# Towards embodied multi-modal video understanding

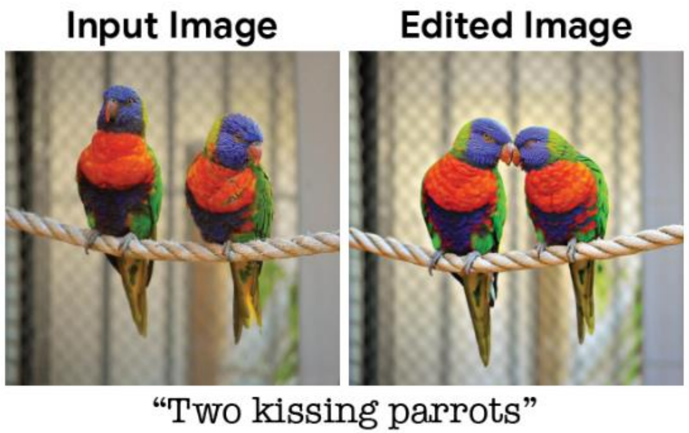

Ivan Laptev

@MBZUAI

# Computer Vision in 2023

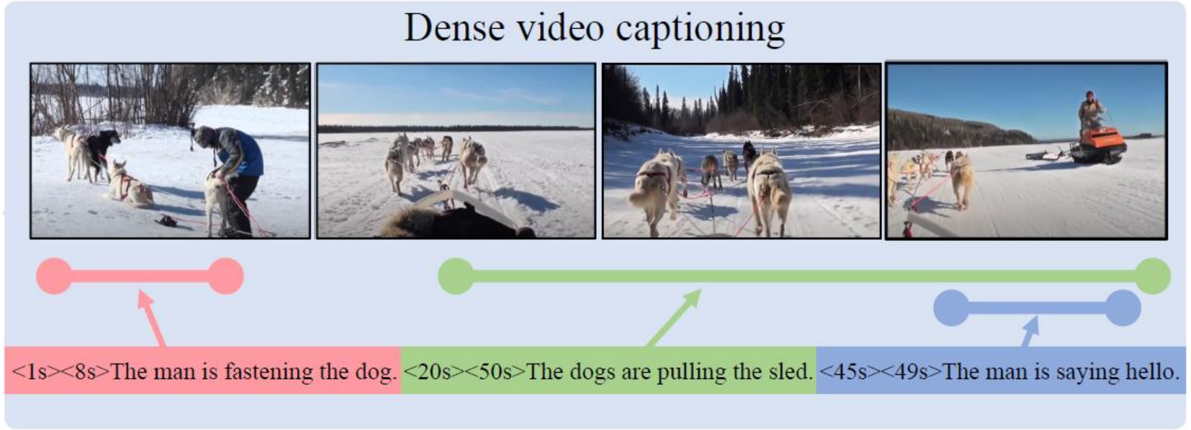



Flamingo: a Visual Language Model for Few-Shot Learning. Alayrac et al., NeurIPS 2022




Imagic: Text-Based Real Image Editing with Diffusion Models. Kawar et al. 2023 arXiv:2210.09276

Reconstructing Hand-Object Interactions in the Wild, Cao et al., CVPR 2021



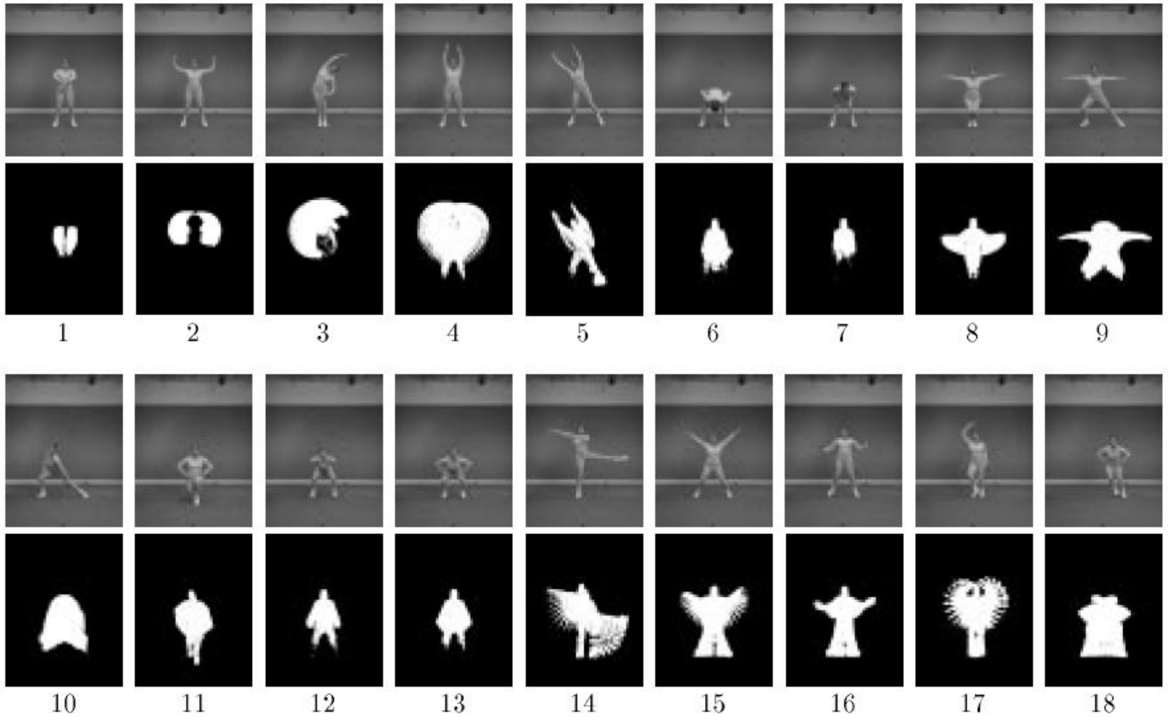

Vid2Seq: Large-Scale Pretraining of a Visual Language Model for Dense Video Captioning, Yang et al., CVPR 2023

# Computer Vision back in 2000

Object Recognition

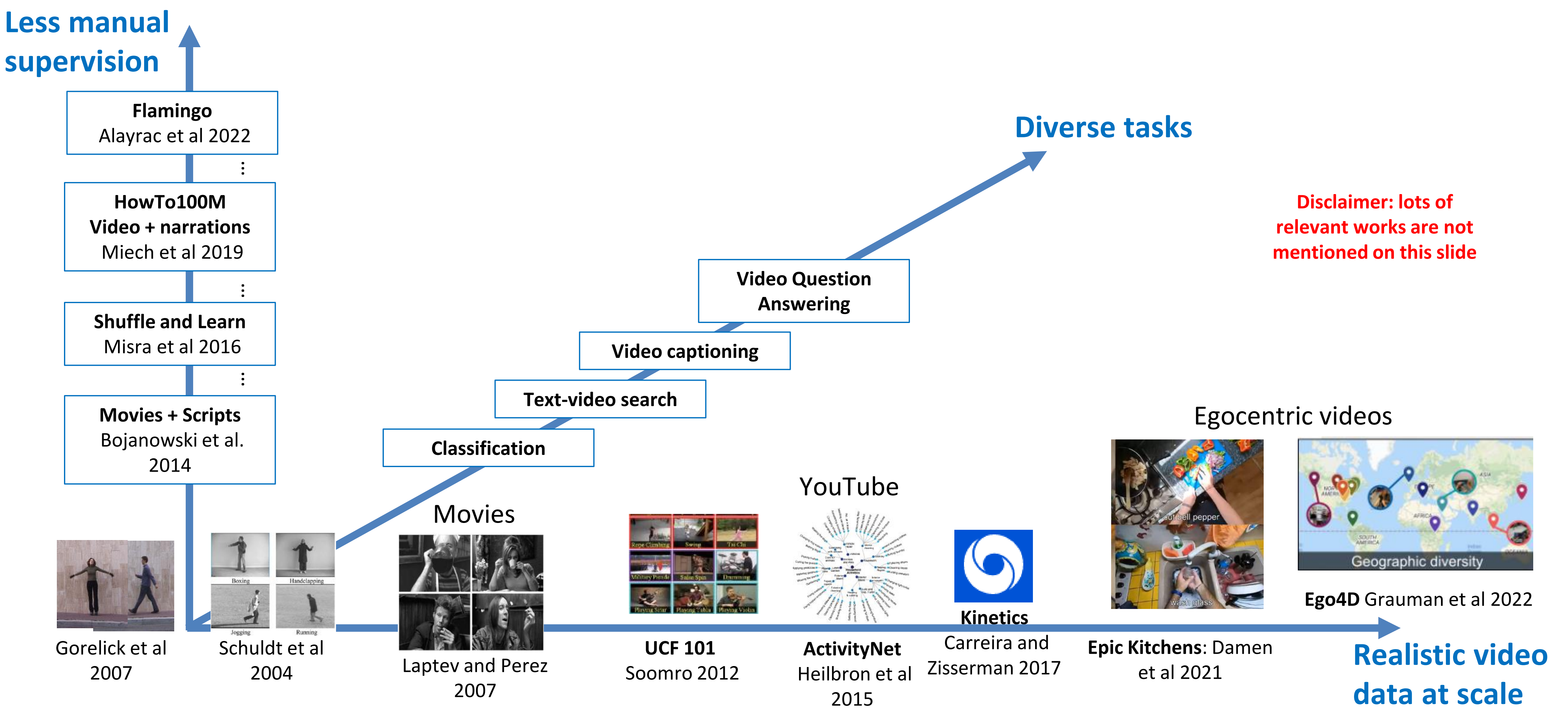

Columbia Object Image Library (COIL-20), Nene et al., 1996

Action recognition




Aerobics dataset: Bobick and Davis, TPAMI 2001

# Video and action recognition in retrospective



# Video and action recognition in retrospective

Less manual supervision

Flamingo
Alayrac et al 2022

HowTo100M
Video + narrations
Miech et al 2019

Shuffle and Learn
Misra et al 2016

Movies + Scripts
Bojanowski et al. 2014

Diverse tasks

Disclaimer: lots of relevant works are not mentioned on this slide

Video Question Answering

Video captioning

Text-video search

Classification

Gorelick et al 2007


Schuldt et al 2004

Movies

Laptev and Perez 2007


**UCF 101**
Soomro 2012

YouTube

**ActivityNet**
Heilbron et al 2015

**Kinetics**
Carreira and Zisserman 2017

Egocentric videos

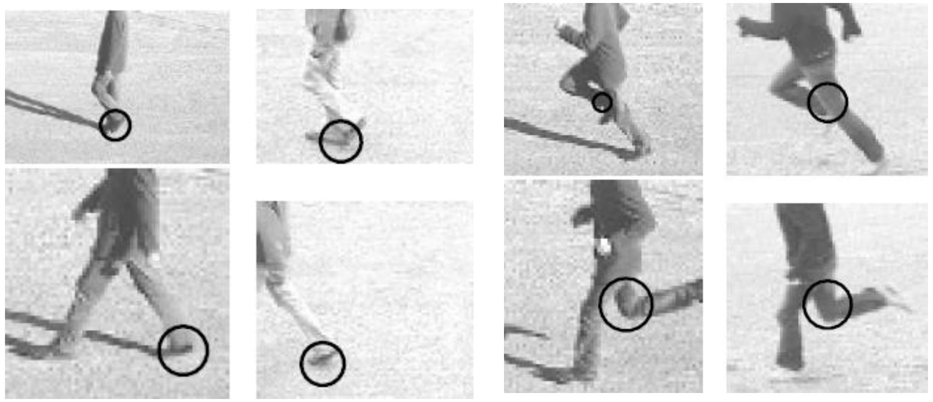

**Epic Kitchens**: Damen et al 2021


**Ego4D** Grauman et al 2022

Realistic video data at scale

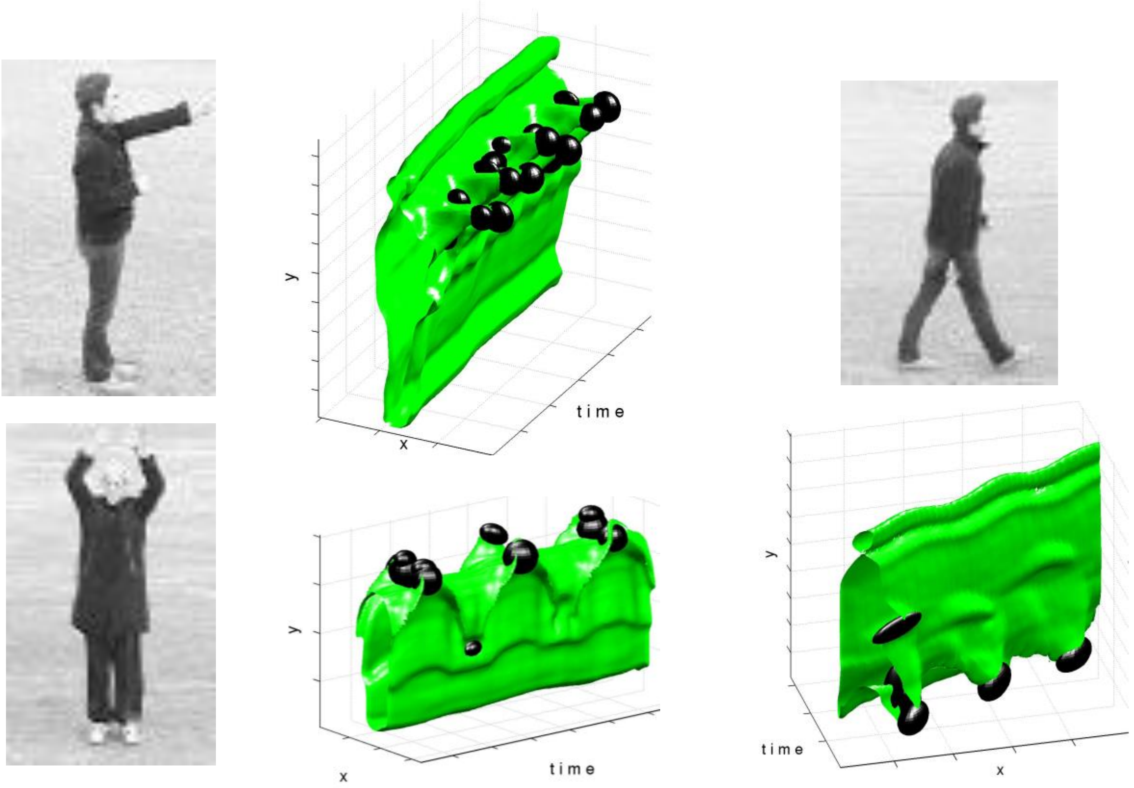
# Representations for video understanding

PhD: Local Spatio-Temporal Image Features for Motion Interpretation
(Laptev 2004, KTH, Stockholm)



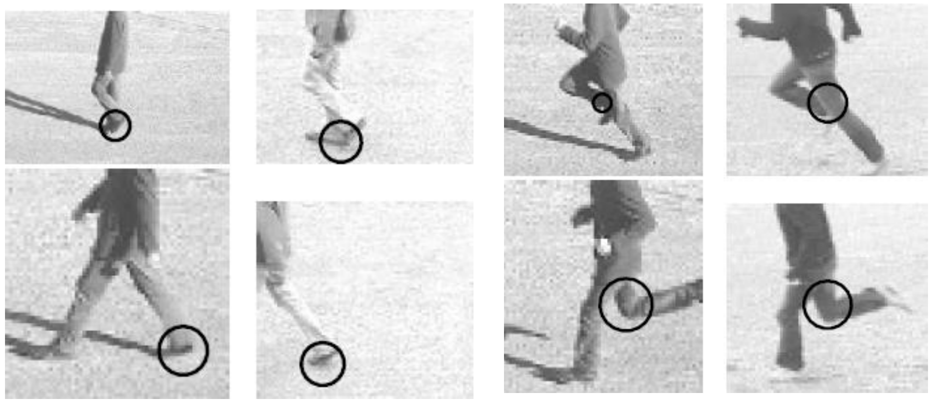
Laptev and Lindeberg ICCV 2003

Space-time interest points



# Representations for video understanding

PhD: Local Spatio-Temporal Image Features for Motion Interpretation
(Laptev 2004, KTH, Stockholm)



Laptev and Lindeberg ICCV 2003

Laptev et al., CVPR 2008, Marszalek et al., CVPR 2009

2017 Helmholtz Prize for fundamental contributions in Computer Vision

# Video and action recognition in retrospective

**Less manual supervision**

**Flamingo**
Alayrac et al 2022

⋮

**HowTo100M**
**Video + narrations**
Miech et al 2019

⋮

**Shuffle and Learn**
Misra et al 2016

⋮

**Movies + Scripts**
Bojanowski et al. 2014

**Diverse tasks**

Disclaimer: lots of relevant works are not mentioned on this slide

**Video Question Answering**

**Video captioning**

**Text-video search**

**Classification**

Gorelick et al 2007


Schuldt et al 2004

Movies



Laptev and Perez 2007



**UCF 101**
Soomro 2012

YouTube



**ActivityNet**
Heilbron et al 2015

**Kinetics**
Carreira and Zisserman 2017

Egocentric videos



**Epic Kitchens**: Damen et al 2021

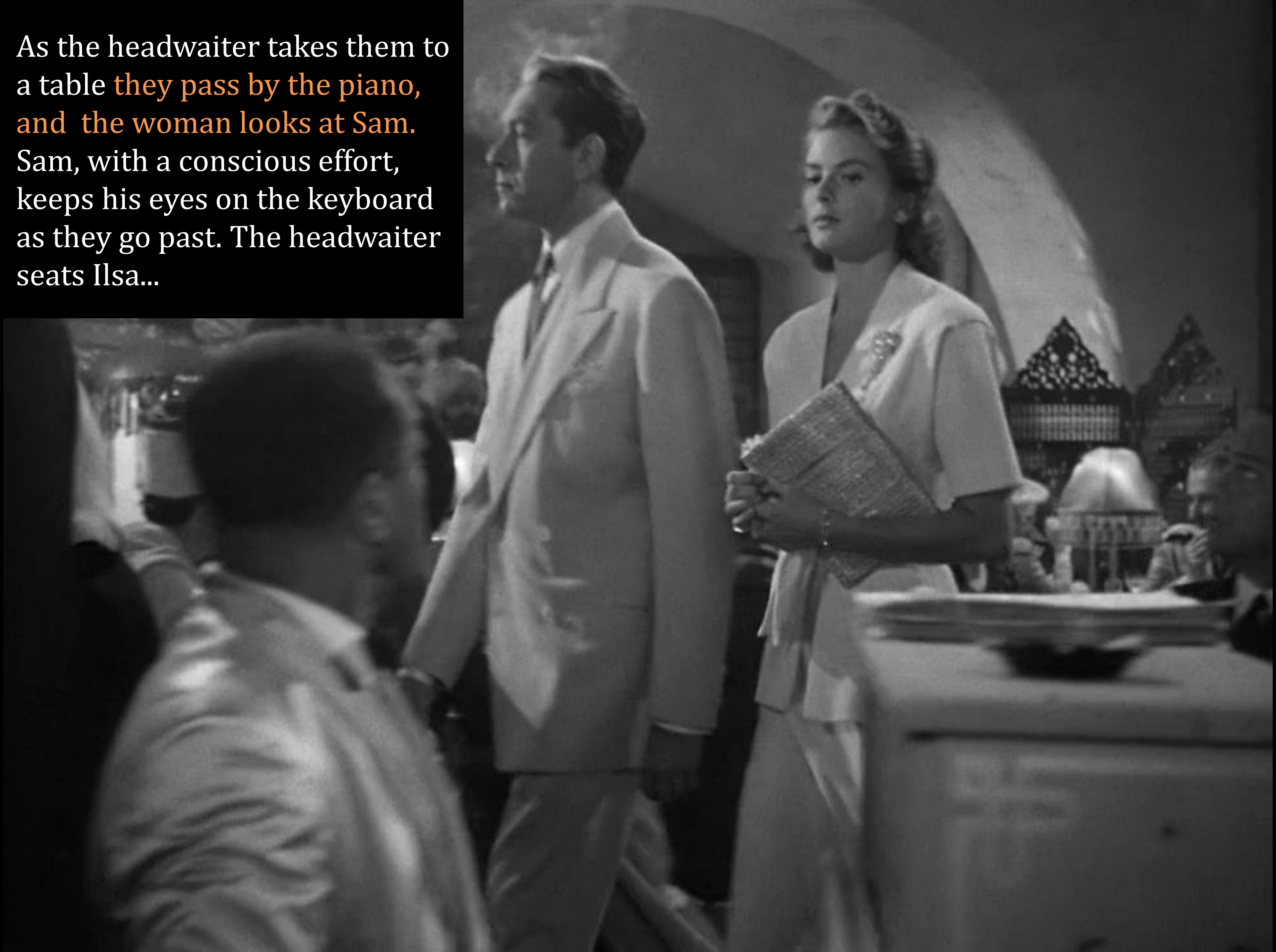



**Ego4D** Grauman et al 2022

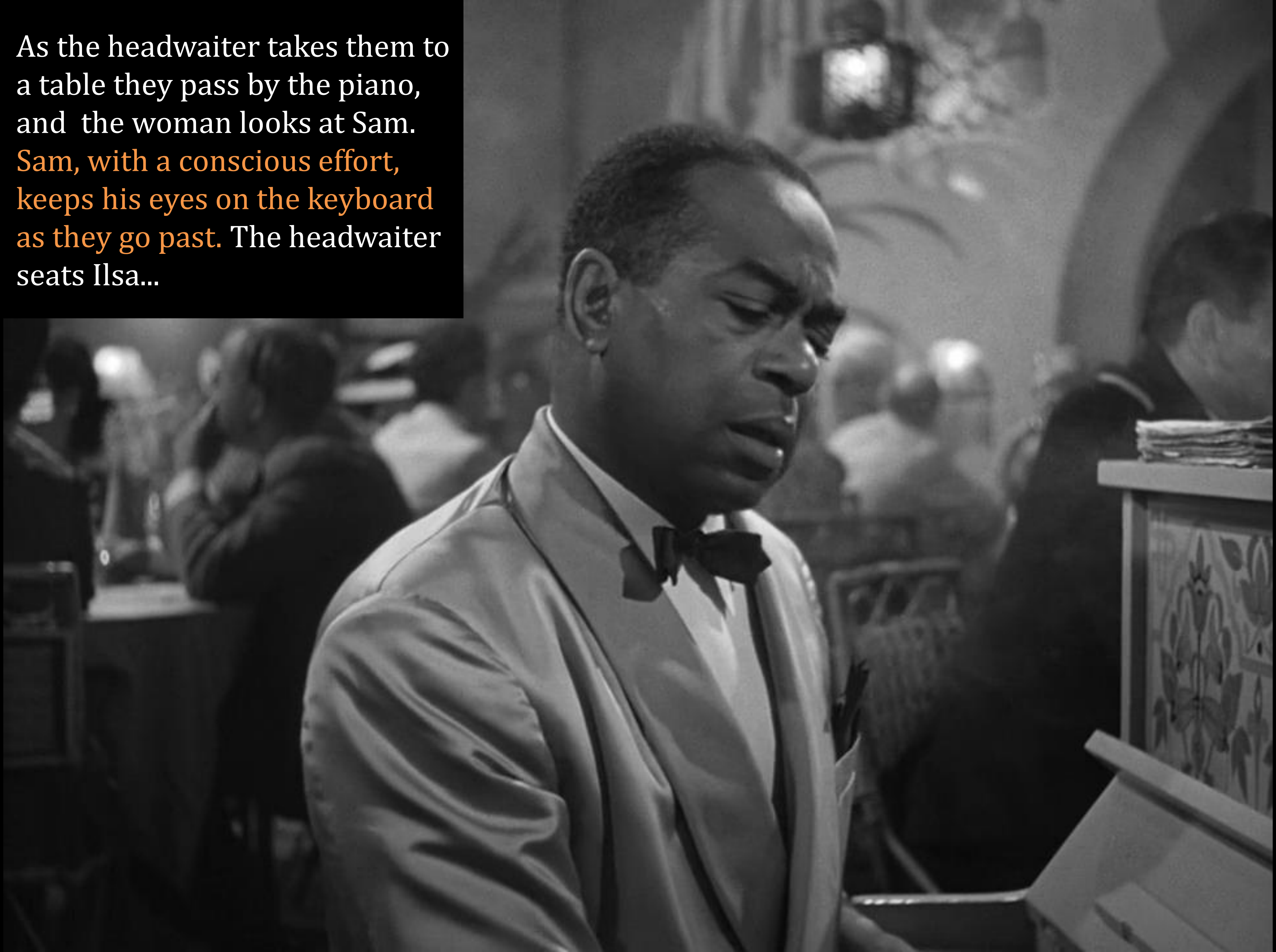
**Realistic video data at scale**

As the headwaiter takes them to a table they pass by the piano, and the woman looks at Sam. Sam, with a conscious effort, keeps his eyes on the keyboard as they go past. The headwaiter seats Ilsa...

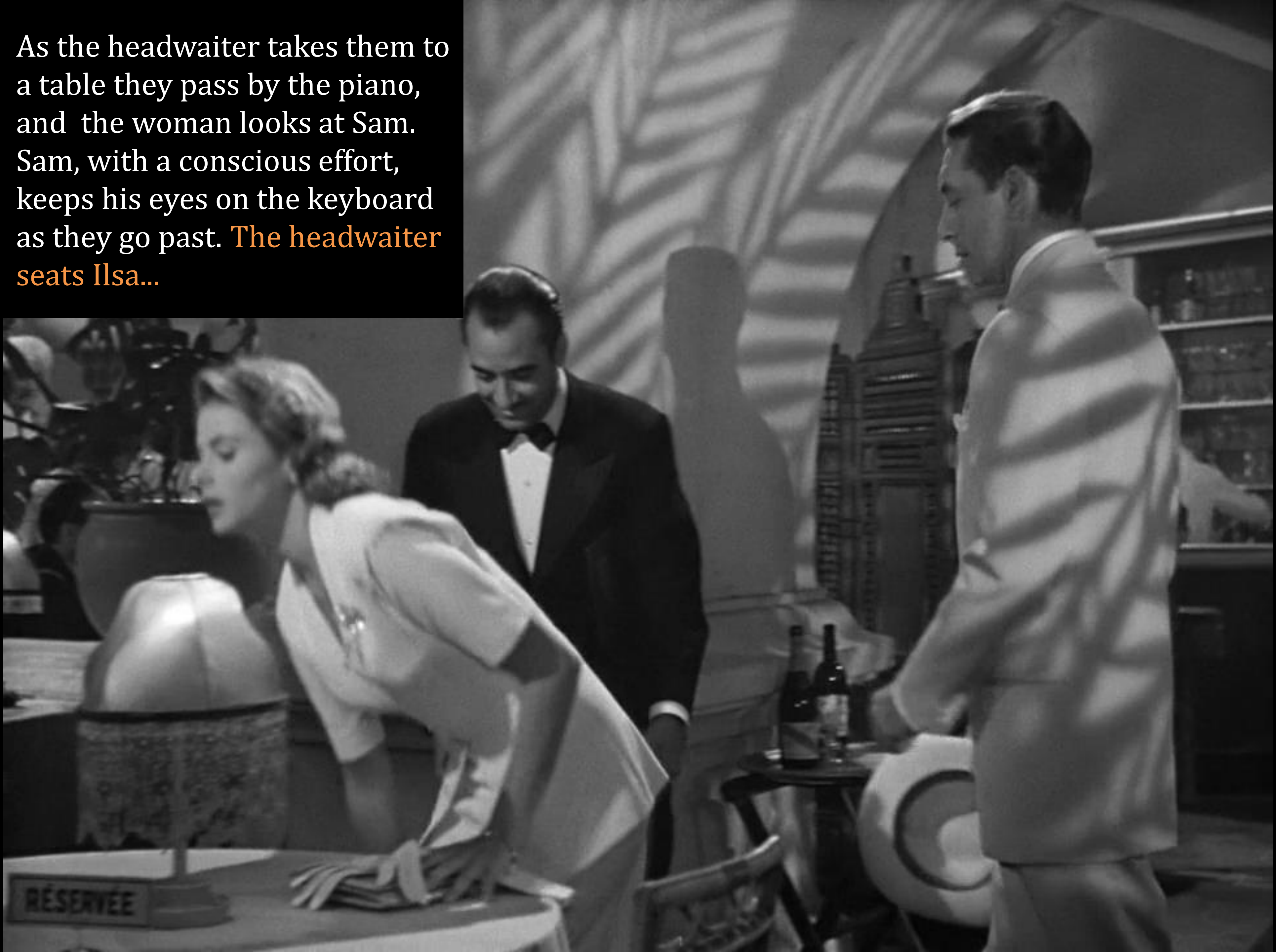
As the headwaiter takes them to a table they pass by the piano, and  the woman looks at Sam. Sam, with a conscious effort, keeps his eyes on the keyboard as they go past. The headwaiter seats Ilsa...



As the headwaiter takes them to a table they pass by the piano, and the woman looks at Sam. Sam, with a conscious effort, keeps his eyes on the keyboard as they go past. The headwaiter seats Ilsa...

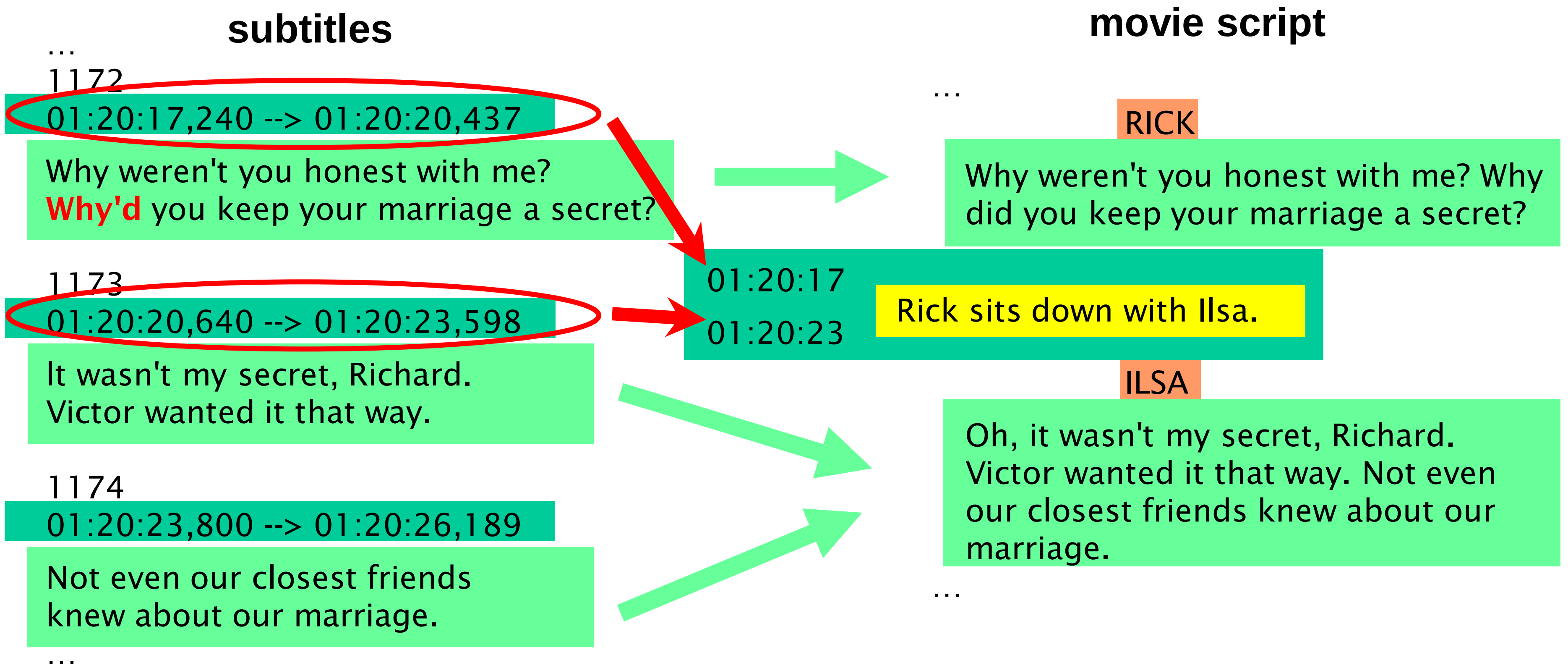


As the headwaiter takes them to a table they pass by the piano, and the woman looks at Sam. Sam, with a conscious effort, keeps his eyes on the keyboard as they go past. The headwaiter seats Ilsa...



# Script-based video annotation

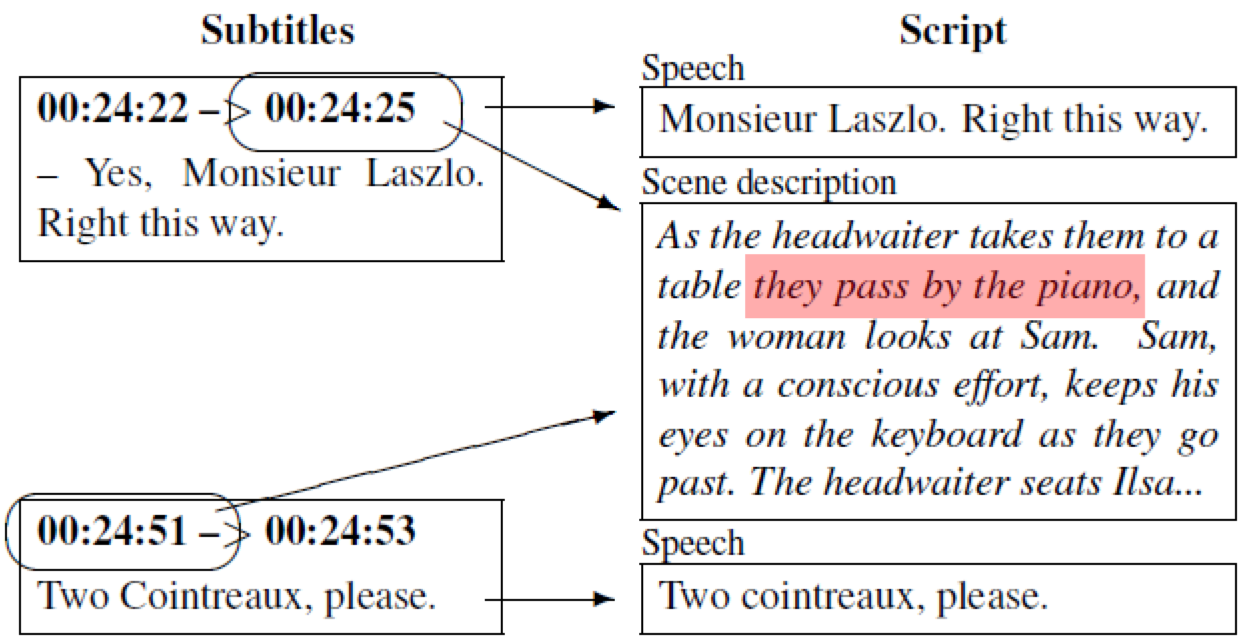
- Scripts available for >500 movies (no time synchronization)
  www.dailyscript.com, www.movie-page.com, www.weeklyscript.com …
- Subtitles (with time info.) are available for the most of movies
- Can transfer time to scripts by text alignment

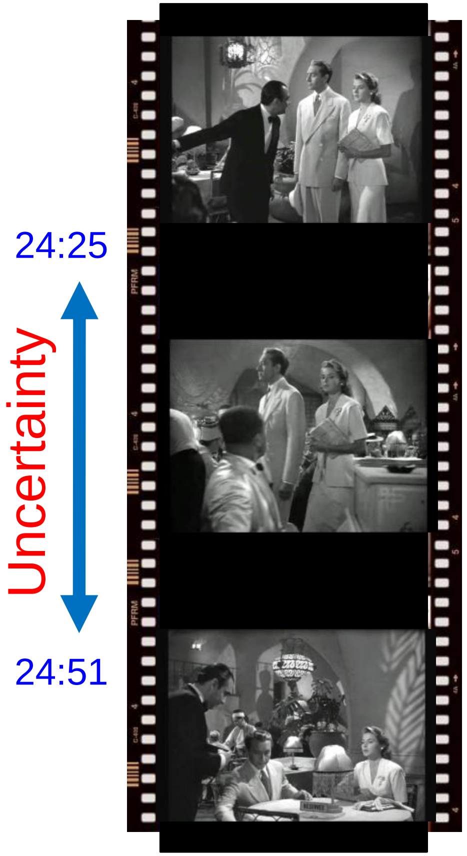


# Scripts as weak supervision

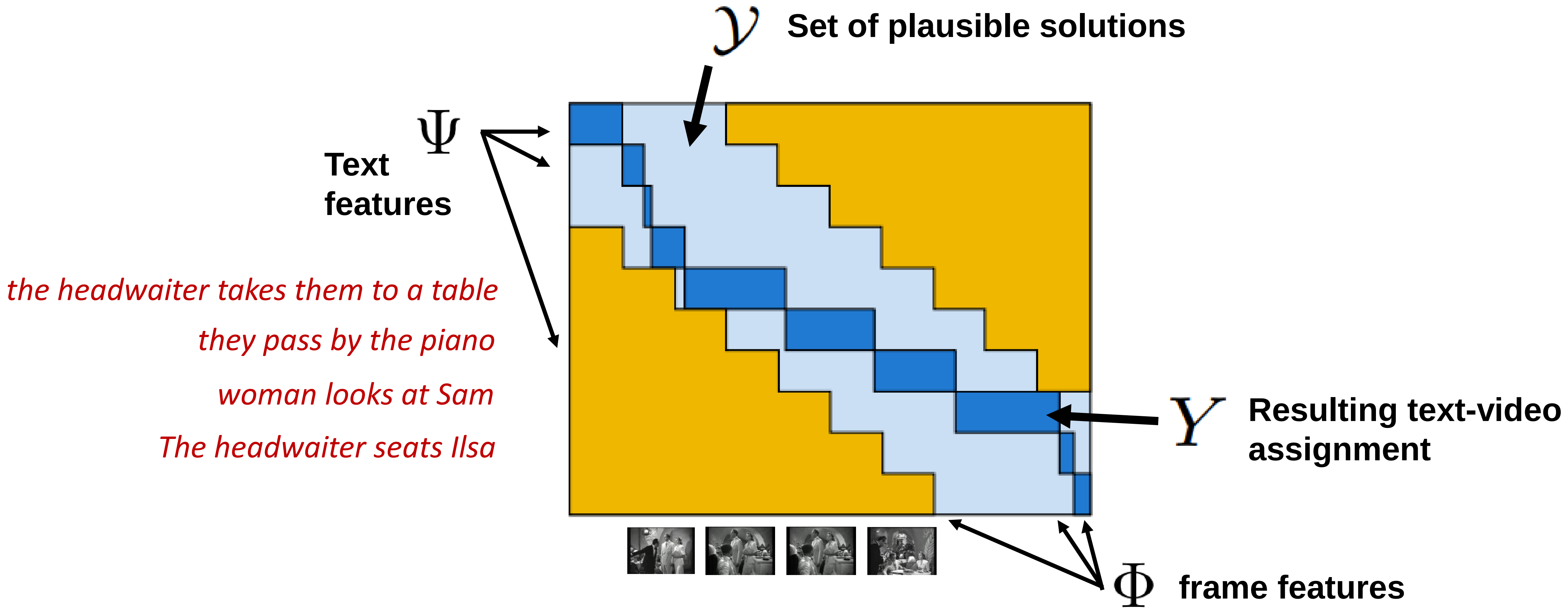
Challenges:

- Imprecise temporal localization
- No explicit spatial localization



# Constrained text-video assignment



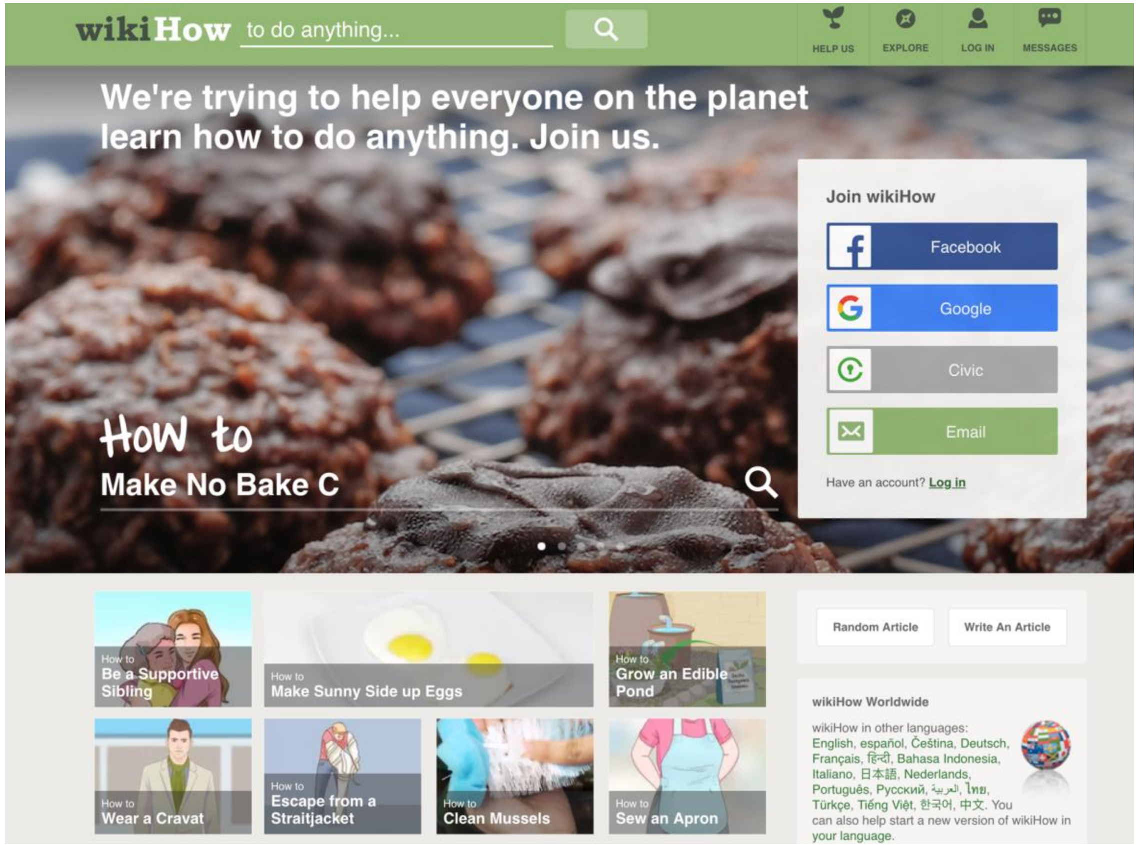

$$\min_{Y \in \mathcal{Y}} \min_{W \in \mathbb{R}^{E \times D}} \frac{1}{2I} \|\Psi Y - W\Phi\|_F^2 + \frac{\lambda}{2} \|W\|_F^2$$

[Bojanowski, Lajugie, Grave, Bach, Laptev, Ponce, Schmid 2015]


ROCKWELL
SPARTANS

# More data: Narrated instructional videos

# Going WikiHow scale




Step 1: Scrap ~130K tasks from WikiHow

**Examples of scrapped tasks**

- ~~How to Be Healthy~~
- How to Cook Quinoa in a Rice Cooker
- How to Sew an Apron
- How to Break a Chain
- ~~How to April Fool your Girlfriend~~
- ....

Step 2: Filter out non-visual tasks

# HowTo100M dataset

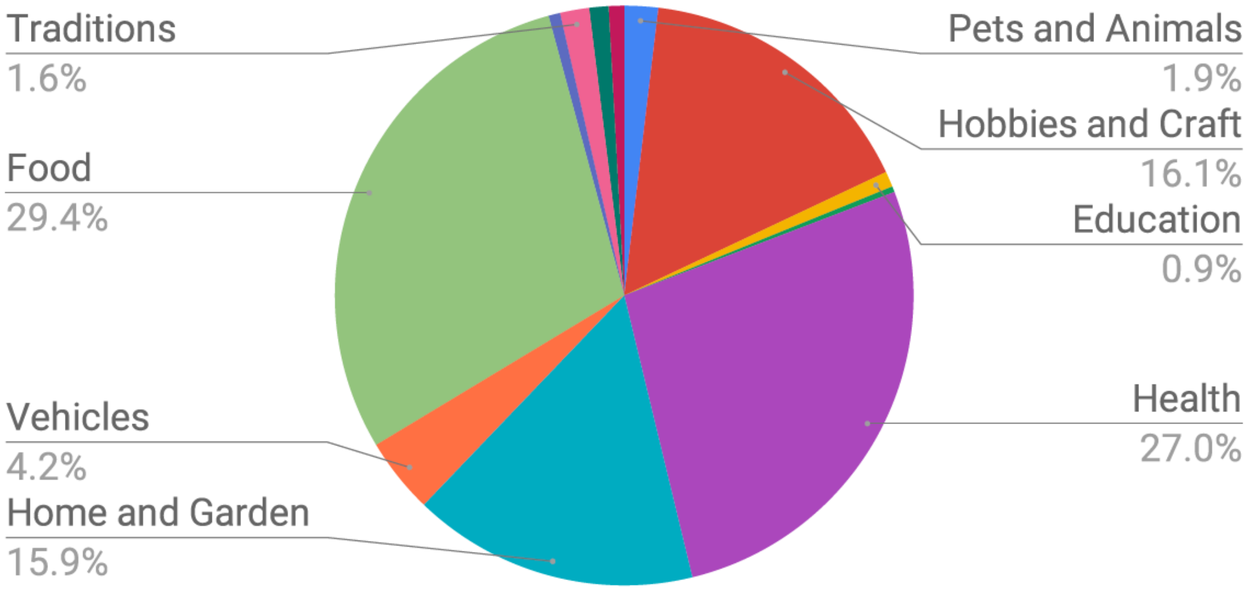
23 k
wikiHow
visual tasks

1.2 M
YouTube
videos

136 M
Short
clips

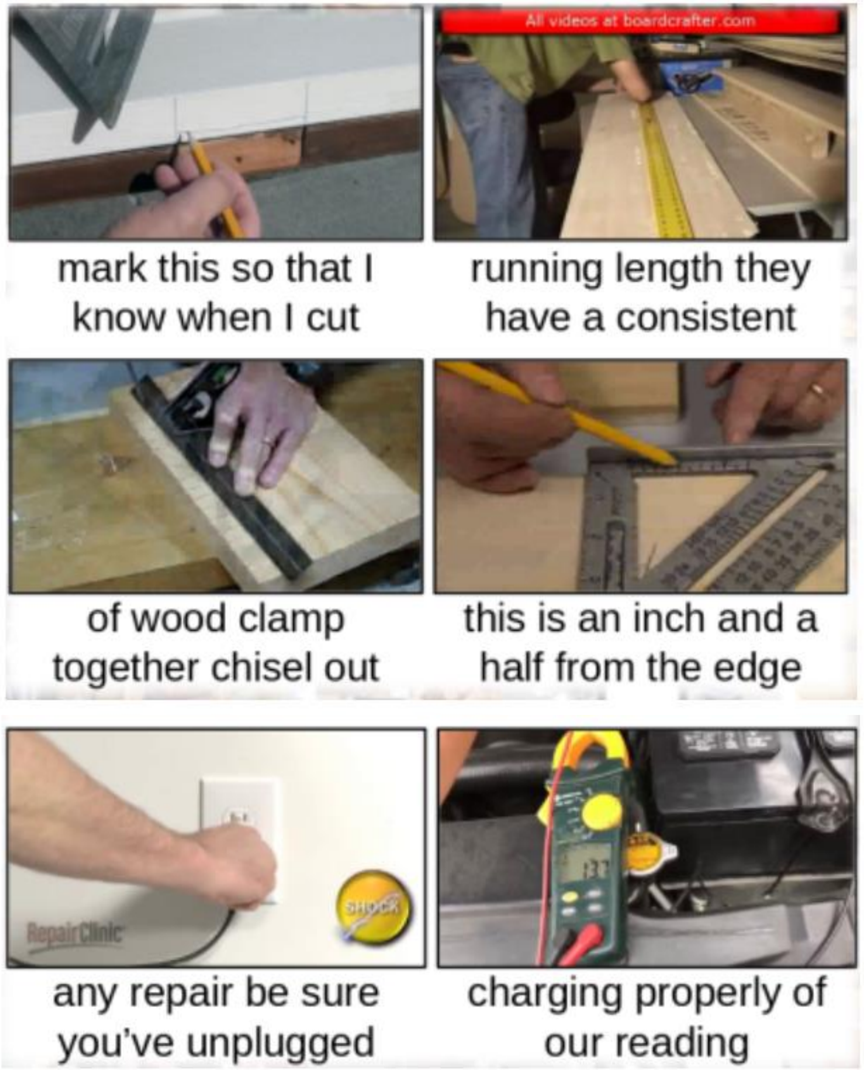
ASR
Supervision
*ASR: Automated Speech Recognition

0
Manual
annotation




[Miech, Zhukov, Alayrac, Tapaswi, Laptev and Sivic, ICCV 2019]

# HowTo100M dataset: Examples

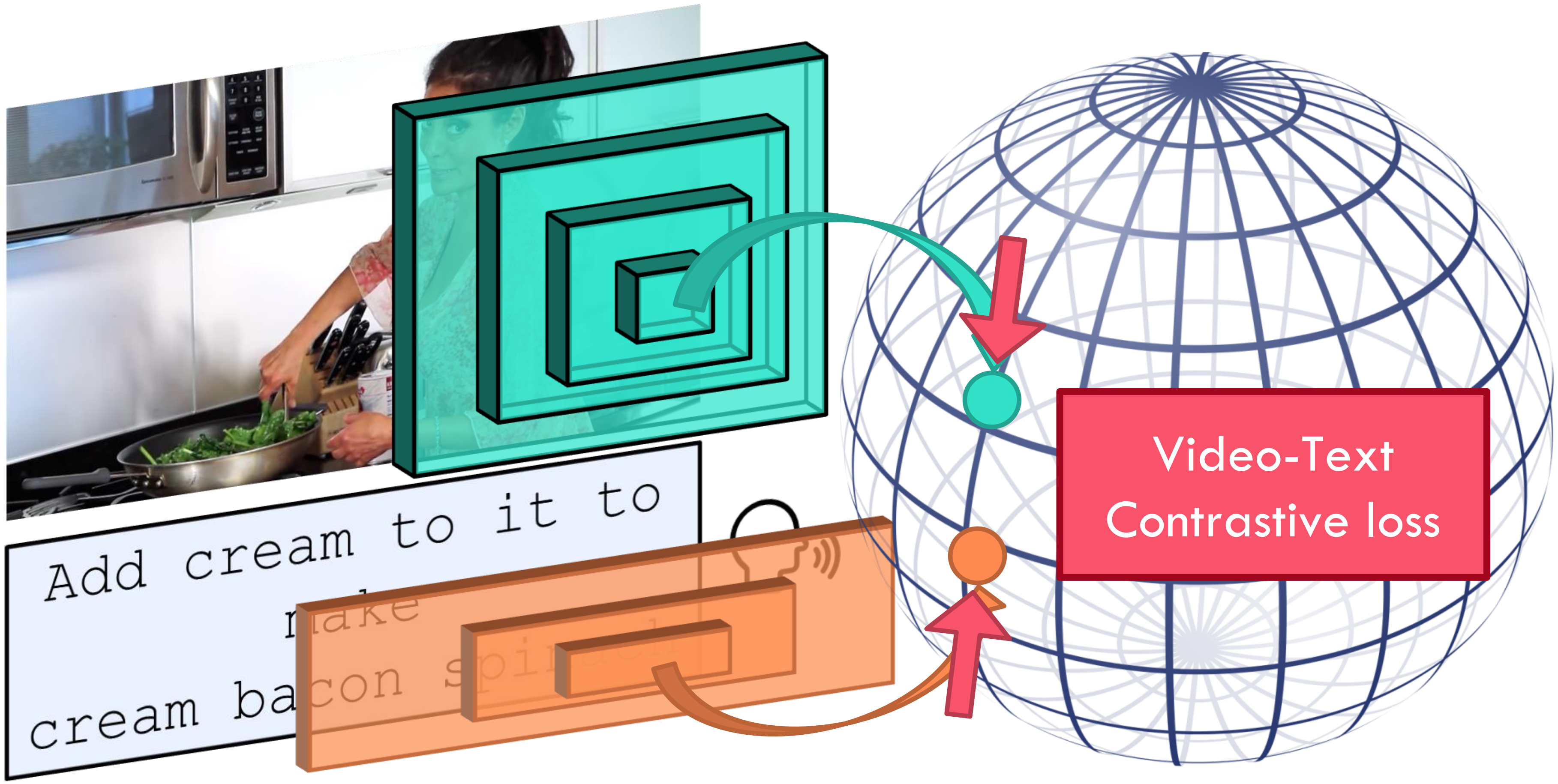

[Miech, Zhukov, Alayrac, Tapaswi, Laptev and Sivic, ICCV 2019]

# Video description datasets

| Dataset | Clips | Captions | Videos | Duration | Source | Year |
|---|---|---|---|---|---|---|
| Charades [42] | 10k | 16k | 10,000 | 82h | Home | 2016 |
| MSR-VTT [52] | 10k | 200k | 7,180 | 40h | Youtube | 2016 |
| YouCook2 [61] | 14k | 14k | 2,000 | 176h | Youtube | 2018 |
| EPIC-KITCHENS [5] | 40k | 40k | 432 | 55h | Home | 2018 |
| DiDeMo [11] | 27k | 41k | 10,464 | 87h | Flickr | 2017 |
| M-VAD [46] | 49k | 56k | 92 | 84h | Movies | 2015 |
| MPII-MD [37] | 69k | 68k | 94 | 41h | Movies | 2015 |
| ANet Captions [22] | 100k | 100k | 20,000 | 849h | Youtube | 2017 |
| TGIF [23] | 102k | 126k | 102,068 | 103h | Tumblr | 2016 |
| LSMDC [38] | 128k | 128k | 200 | 150h | Movies | 2017 |
| How2 [39] | 185k | 185k | 13,168 | 298h | Youtube | 2018 |
| **HowTo100M** | **136M** | **136M** | **1.221M** | **134,472h** | Youtube | 2019 |

[Miech, Zhukov, Alayrac, Tapaswi, Laptev and Sivic, ICCV 2019]

# Learning joint text-video embedding




[Miech, Alayrac, Laptev, Smaira, Sivic and Zisserman, CVPR 2020]


[Miech, Alayrac, Laptev, Smaira, Sivic and Zisserman, CVPR 2020]

Time

spinachs what's the name

keep it simple you just want to add

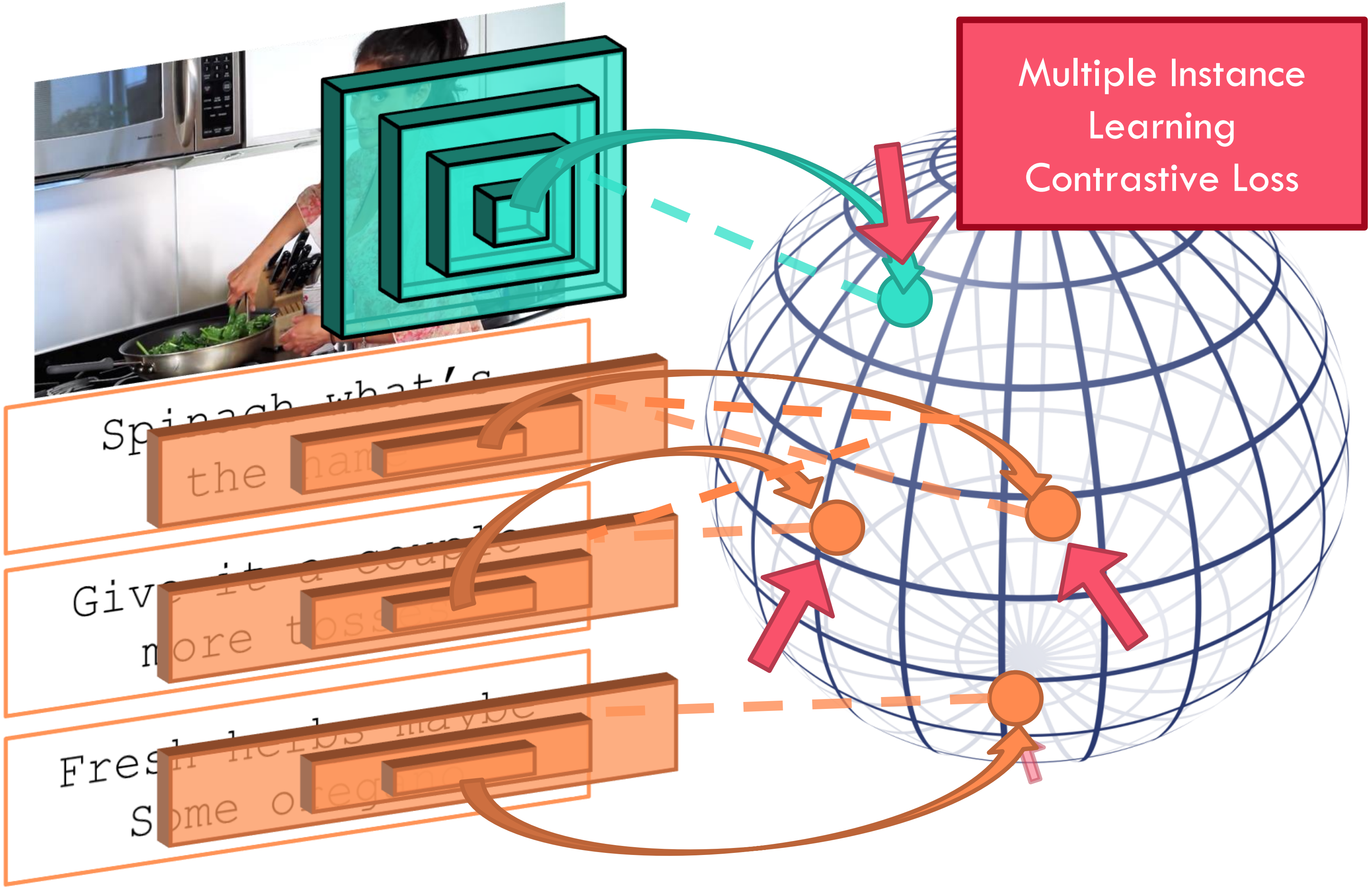
fresh herbs maybe some oregano

you can add cilantro basil they give

give it a couple more tosses

[Miech, Alayrac, Laptev, Smaira, Sivic and Zisserman, CVPR 2020]

# Learning joint text-video embedding




[Miech, Alayrac, Laptev, Smaira, Sivic and Zisserman, CVPR 2020]

# Our formulation: MIL-NCE

Video input

Narration input

$$e^{f(x)^\top g(y)}$$

$f, g$

Video network

Text network

Bag of positive candidate pairs

Negative video-narration pairs

[Miech, Alayrac, Laptev, Smaira, Sivic and Zisserman, CVPR 2020]

# Our formulation: MIL-NCE


$$\max_{f,g} \sum_{i=1}^{n} \log\left(\frac{\sum_{(x,y)\in\mathcal{P}_i} e^{f(x)^\top g(y)}}{\sum_{(x,y)\in\mathcal{P}_i} e^{f(x)^\top g(y)} + \sum_{(x',y')\sim\mathcal{N}_i} e^{f(x')^\top g(y')}}\right)$$

Bag of positive candidate pairs

[Miech, Alayrac, Laptev, Smaira, Sivic and Zisserman, CVPR 2020]

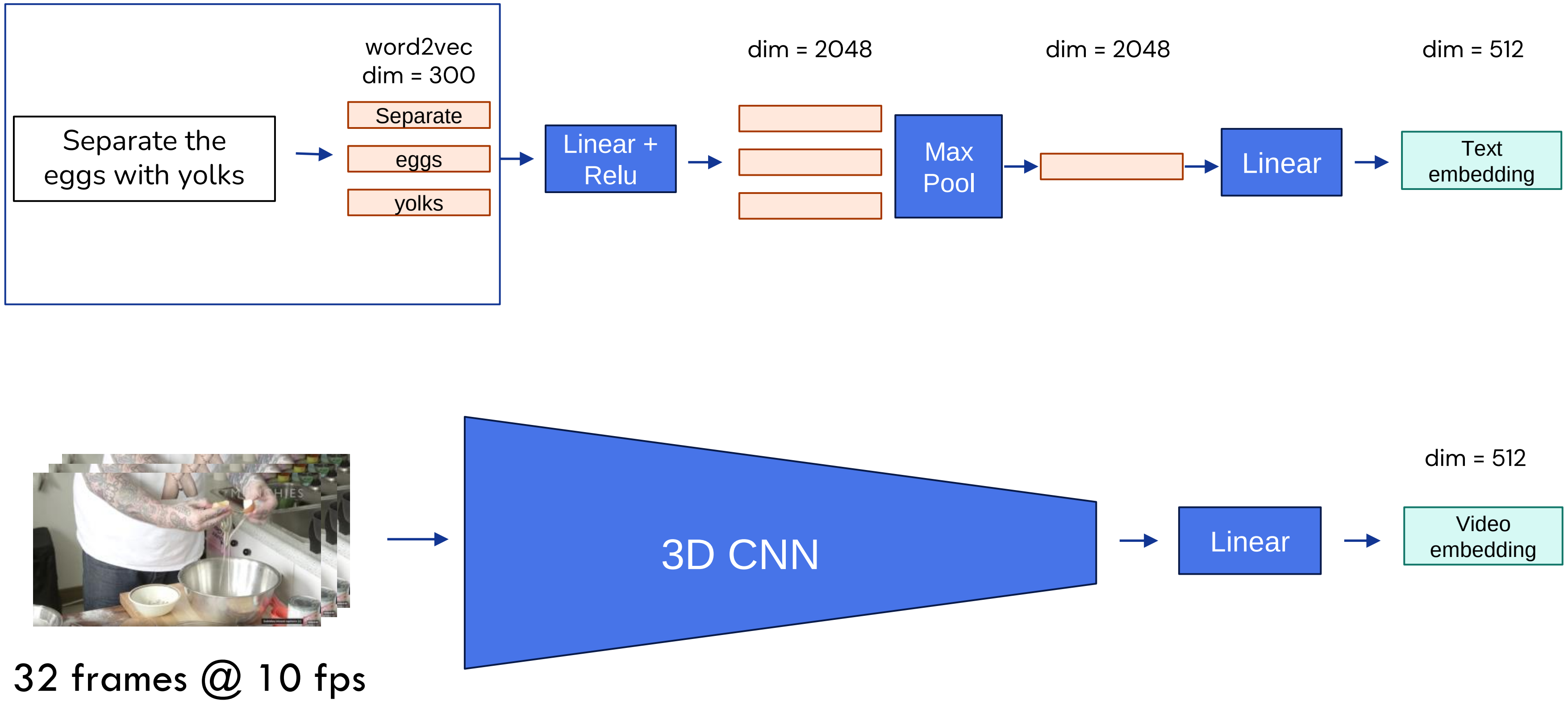
# Our formulation: MIL-NCE


$$\max_{f,g} \sum_{i=1}^{n} \log \left( \frac{\sum\limits_{(x,y)\in\mathcal{P}_i} e^{f(x)^\top g(y)}}{\sum\limits_{(x,y)\in\mathcal{P}_i} e^{f(x)^\top g(y)} + \sum\limits_{(x',y')\sim\mathcal{N}_i} e^{f(x')^\top g(y')}} \right)$$

Negative
video-narration pairs

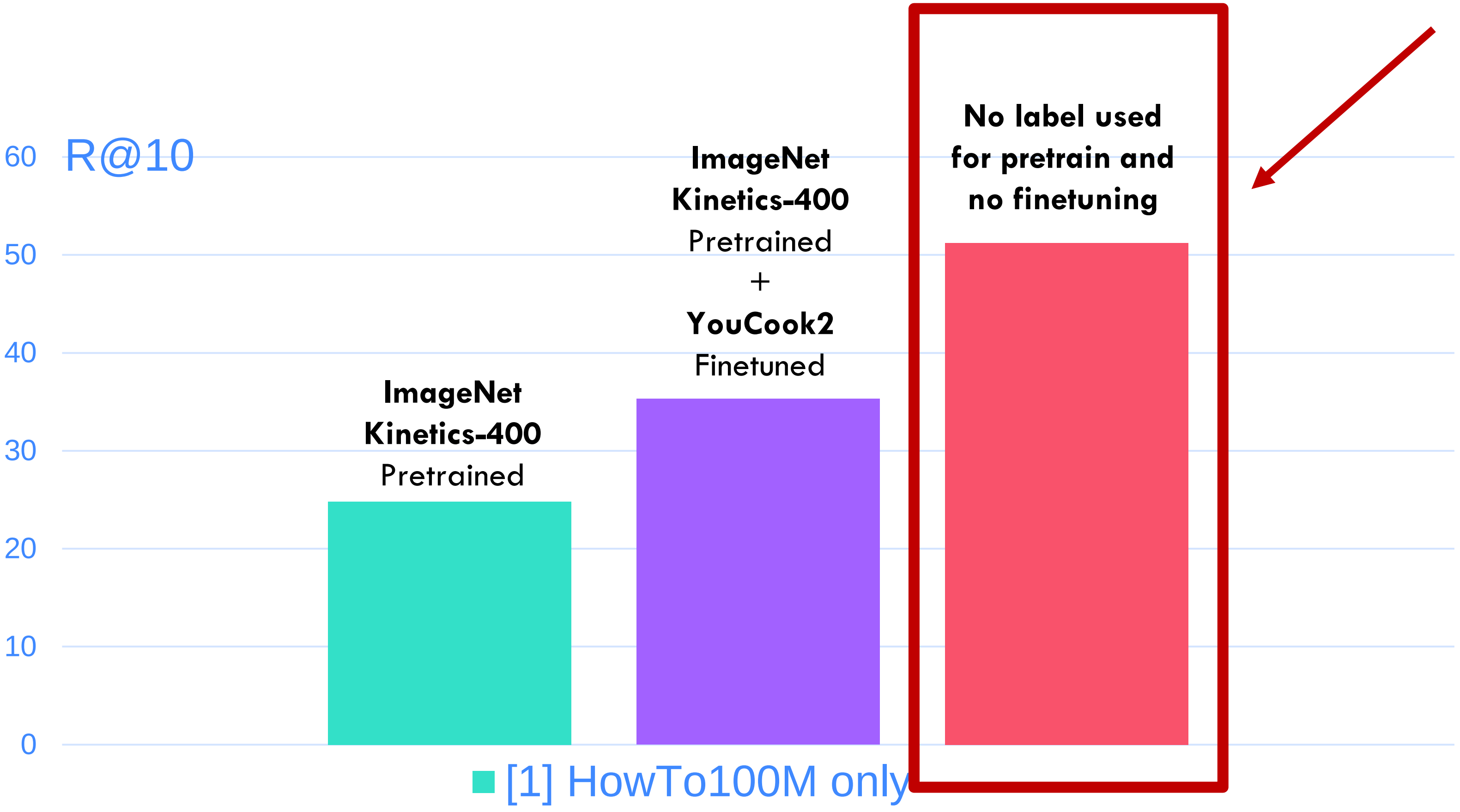
[Miech, Alayrac, Laptev, Smaira, Sivic and Zisserman, CVPR 2020]

# Video-Text model architecture




[Miech, Alayrac, Laptev, Smaira, Sivic and Zisserman, CVPR 2020]

# YouCook2 Zero-Shot Text-to-Video retrieval

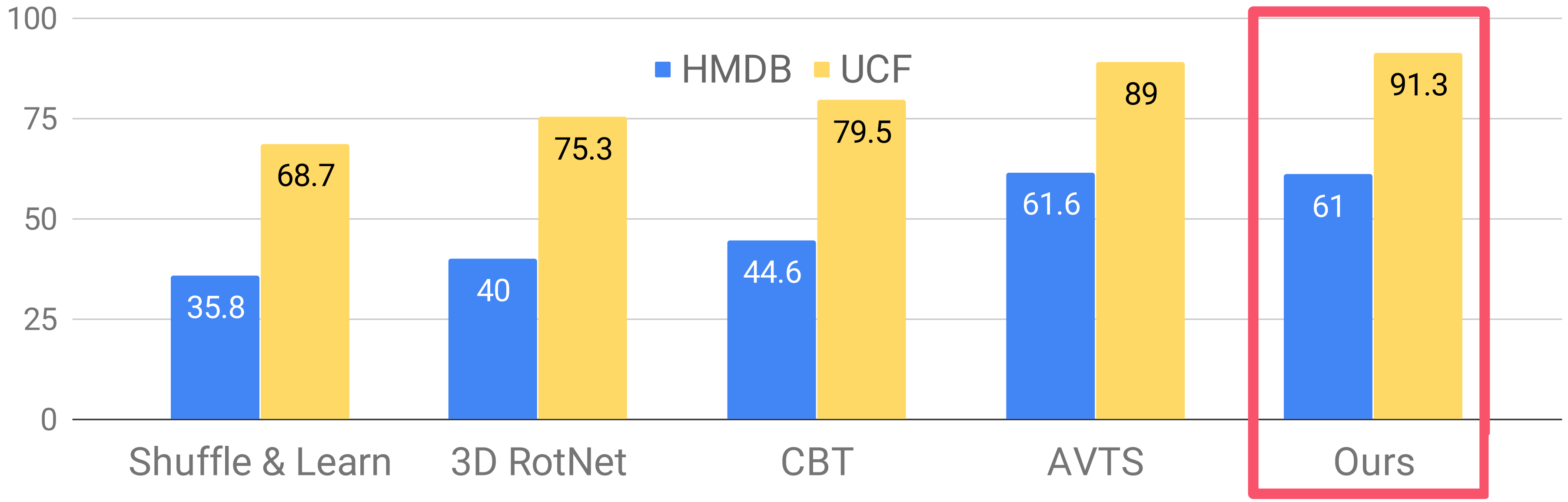



First time a method trained with no manual supervision beats fully-supervised methods

[1] *A.* Miech, D. Zhukov, J.-B. Alayrac, M. Tapaswi, I. Laptev, J. Sivic,
*HowTo100M: Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips*, in ICCV, 2019.

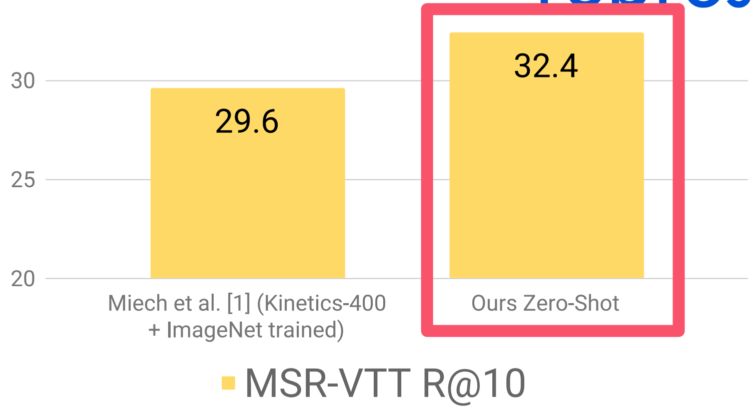
[Miech, Alayrac, Laptev, Smaira, Sivic and Zisserman, CVPR 2020]

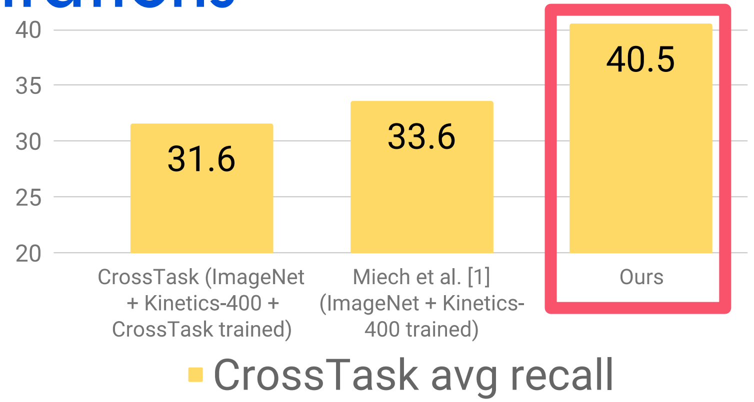
# Action recognition: comparison to self-supervised video representations

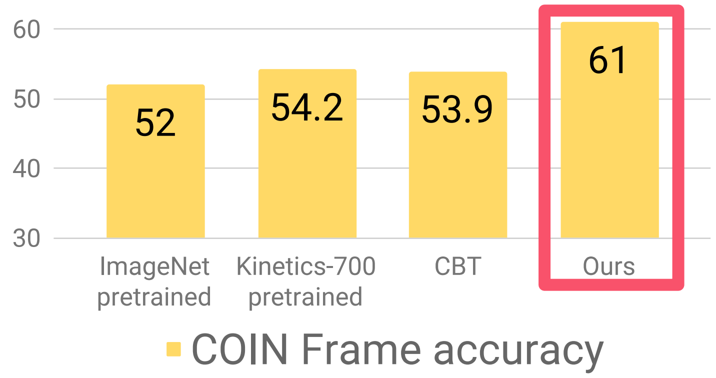


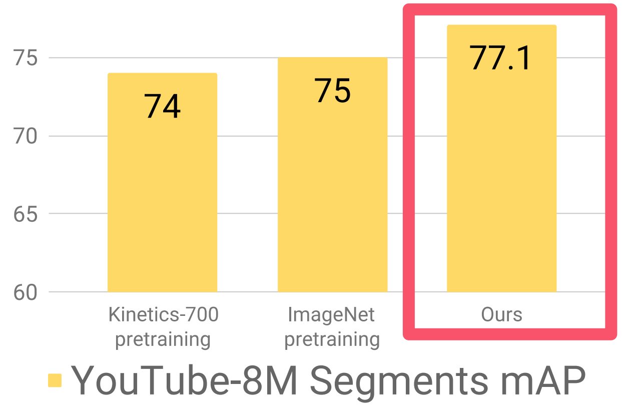

[Miech, Alayrac, Laptev, Smaira, Sivic and Zisserman, CVPR 2020]

# Comparison to fully-supervised representations



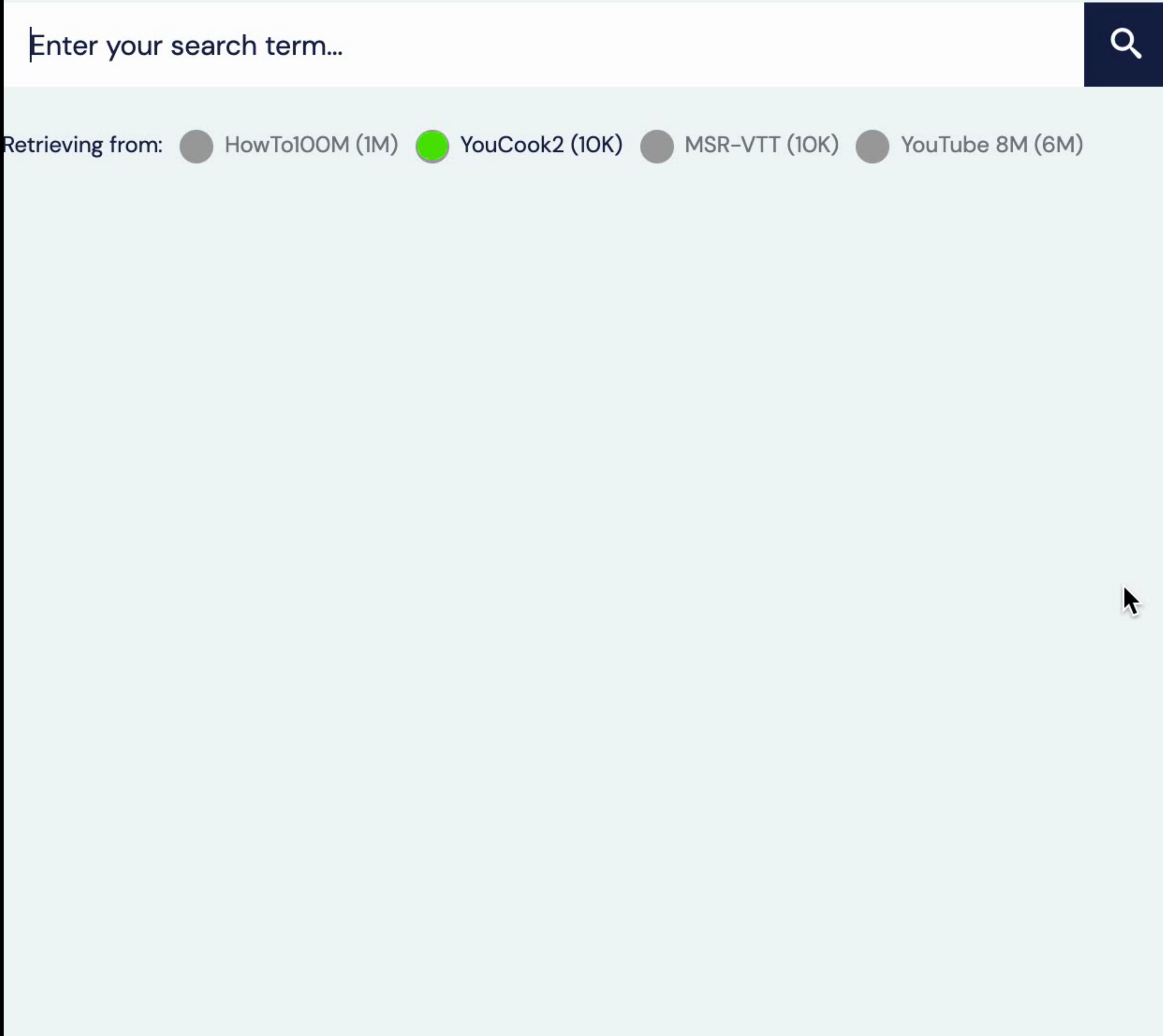




[1] *A.* Miech, D. Zhukov, J.-B. Alayrac, M. Tapaswi, I. Laptev, J. Sivic, *HowTo100M: Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips*, in ICCV, 2019.

[Miech, Alayrac, Laptev, Smaira, Sivic and Zisserman, CVPR 2020]

# Video search by text

https://howto100m.inria.fr



# Video and action recognition in retrospective

Less manual supervision

**Flamingo**
Alayrac et al 2022

**HowTo100M**
**Video + narrations**
Miech et al 2019

**Shuffle and Learn**
Misra et al 2016

**Movies + Scripts**
Bojanowski et al. 2014

**Diverse tasks**

Disclaimer: lots of relevant works are not mentioned on this slide

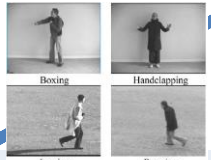
**Video Question Answering**

**Video captioning**

**Text-video search**

**Classification**

Gorelick et al 2007




Schuldt et al 2004

Movies



Laptev and Perez 2007



**UCF 101**
Soomro 2012

YouTube



**ActivityNet**
Heilbron et al 2015

**Kinetics**
Carreira and Zisserman 2017

Egocentric videos



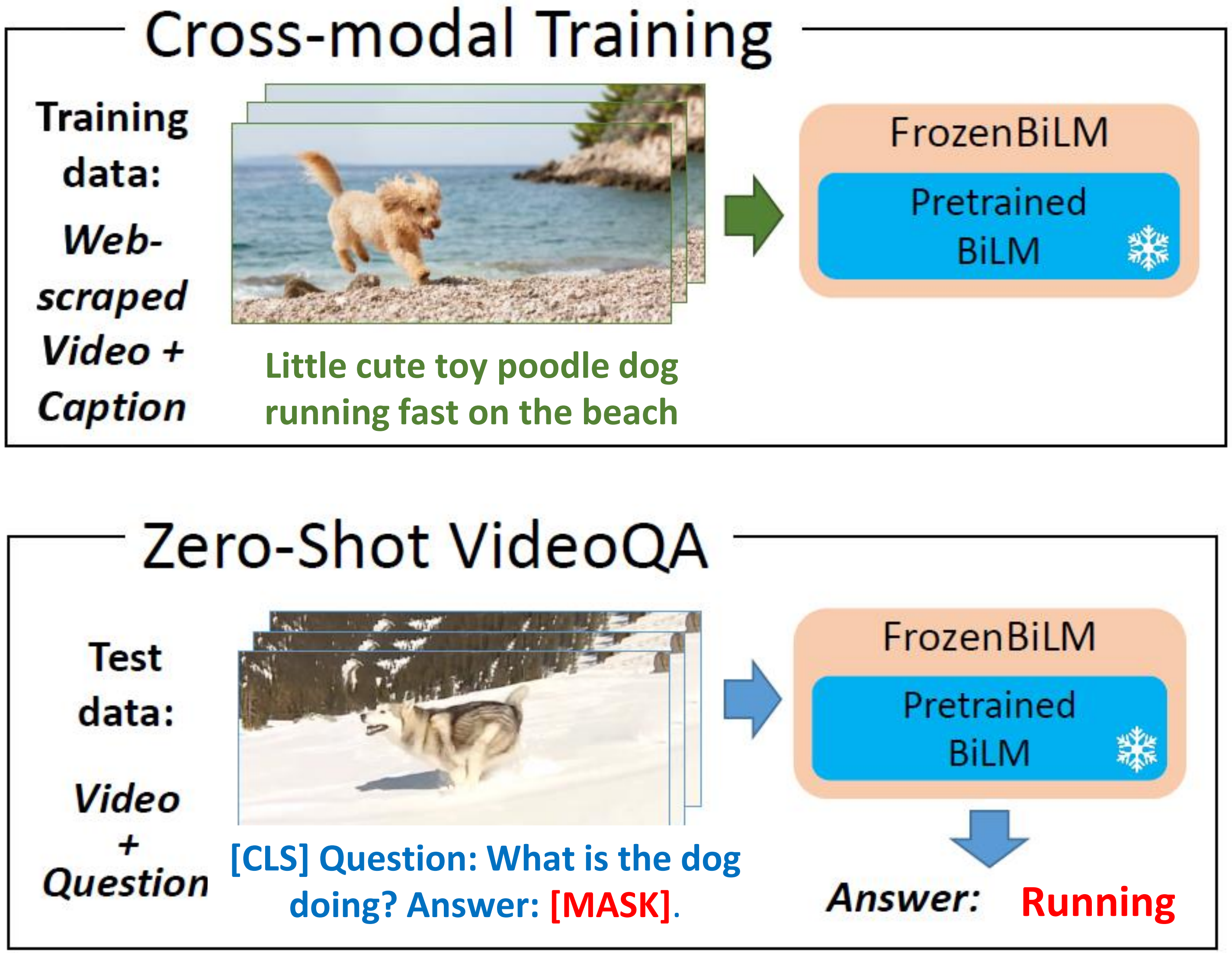
**Epic Kitchens**: Damen et al 2021




**Ego4D** Grauman et al 2022

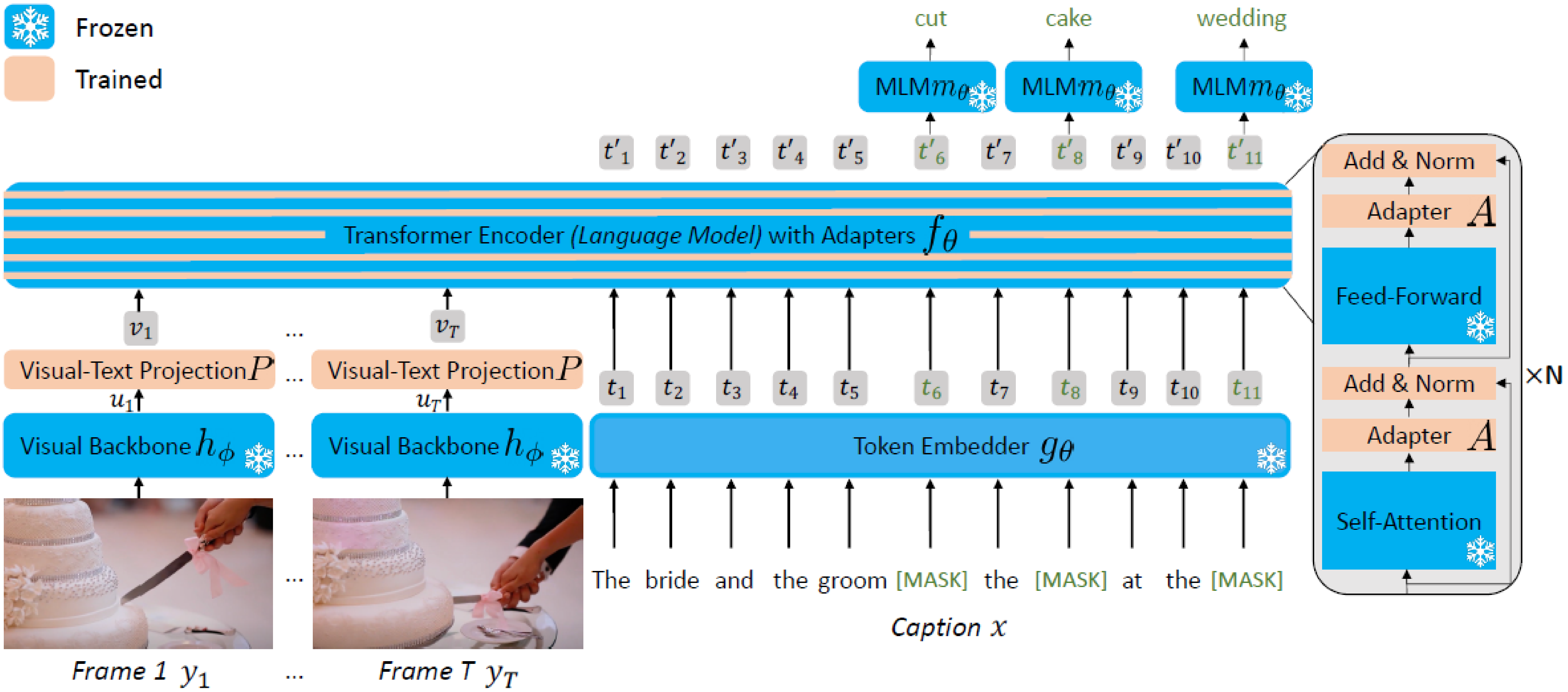
**Realistic video data at scale**

# Zero-Shot Video Question Answering



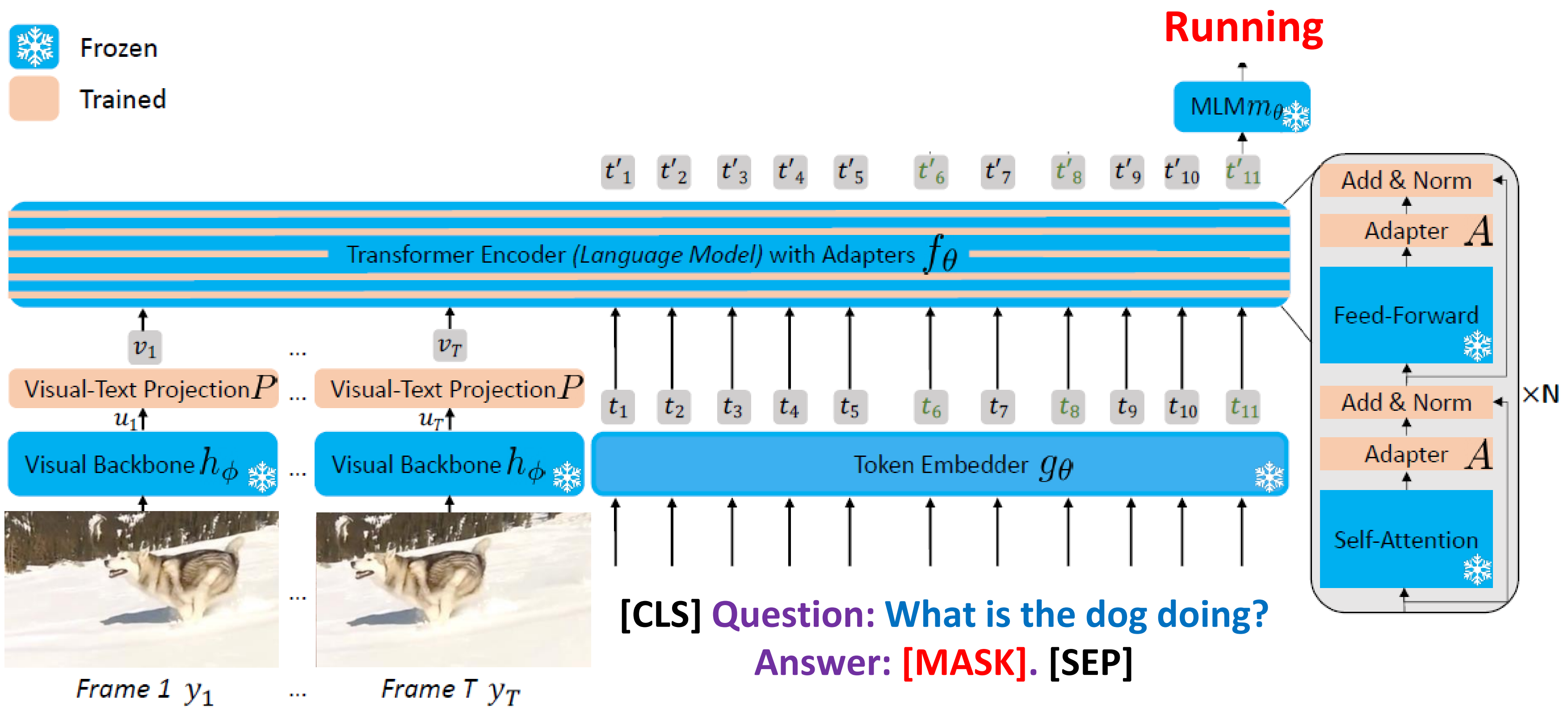

[Yang, Miech, Sivic, Laptev and Schmid, NeurIPS 2022]

# FrozenBiLM: Training




[Yang, Miech, Sivic, Laptev and Schmid, NeurIPS 2022]

# FrozenBiLM: Zero-Shot VideoQA

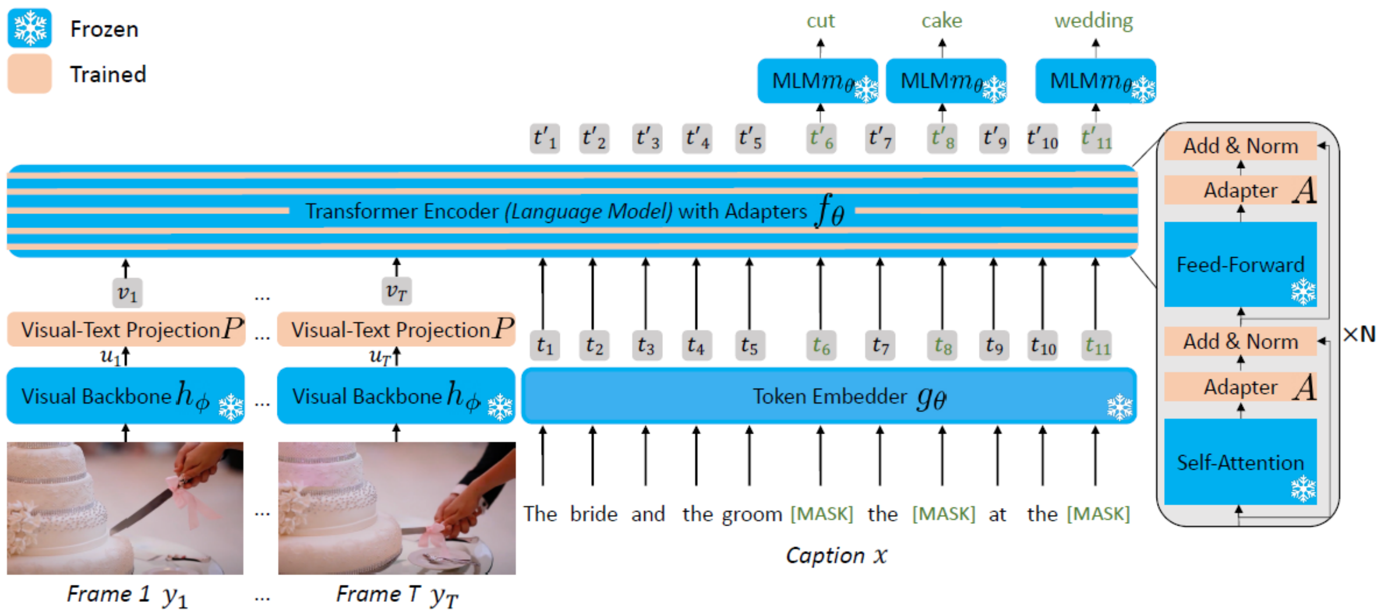



**Input prompt engineering**

| | |
|---|---|
| *Open-ended VideoQA* | "[CLS] **Question:** <Question>? **Answer:** [MASK]. [SEP]" |
| *Multiple-choice VideoQA* | "[CLS] **Question:** <Question>? **Is it** '<Answer Candidate>'? [MASK]. [SEP]" |
| *Video-conditioned fill-in-the-blank task* | "[CLS] <Sentence with a [MASK] token>. [SEP]" |

[Yang, Miech, Sivic, Laptev and Schmid, NeurIPS 2022]

# FrozenBiLM: Zero-Shot SOTA comparison

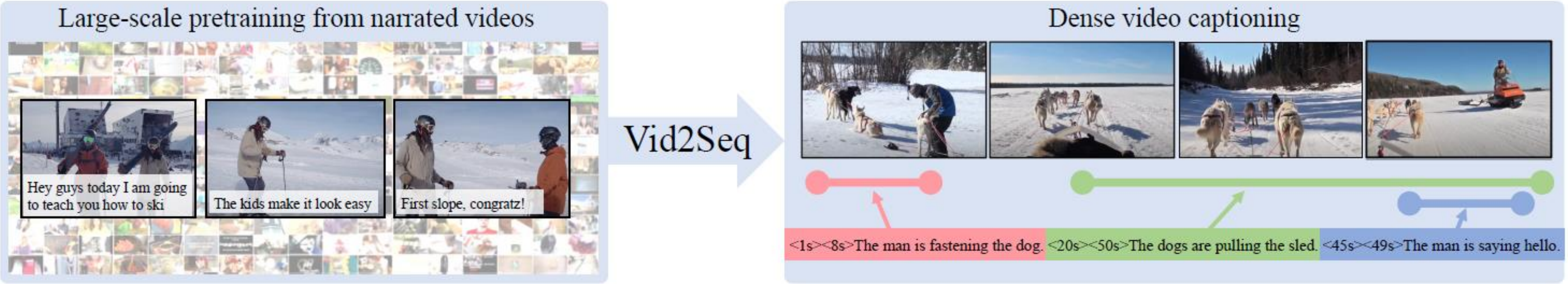



| Method | Training Data | Fill-in-the-blank LSMDC | Open-ended iVQA | MSRVTT-QA | MSVD-QA | ActivityNet-QA | TGIF-QA | Multiple-choice How2QA | TVQA |
|---|---|---|---|---|---|---|---|---|---|
| Random | — | 0.1 | 0.1 | 0.1 | 0.1 | 0.1 | 0.1 | 25 | 20 |
| CLIP ViT-L/14 [75] | 400M image-texts | 1.2 | 9.2 | 2.1 | 7.2 | 1.2 | 3.6 | 47.7 | 26.1 |
| Just Ask [108] | HowToVQA69M + WebVidVQA3M | — | 13.3 | 5.6 | 13.5 | 12.3 | — | 53.1 | — |
| Reserve [116] | YT-Temporal-1B | 31.0 | — | 5.8 | — | — | — | — | — |
| *FrozenBiLM* (Ours) | WebVid10M | **51.5** | **26.8** | **16.7** | **33.8** | **25.9** | **41.9** | **58.4** | **59.7** |

[Yang, Miech, Sivic, Laptev and Schmid, NeurIPS 2022]

# Vid2Seq: Large-Scale Pretraining of a Visual Language Model for Dense Video Captioning

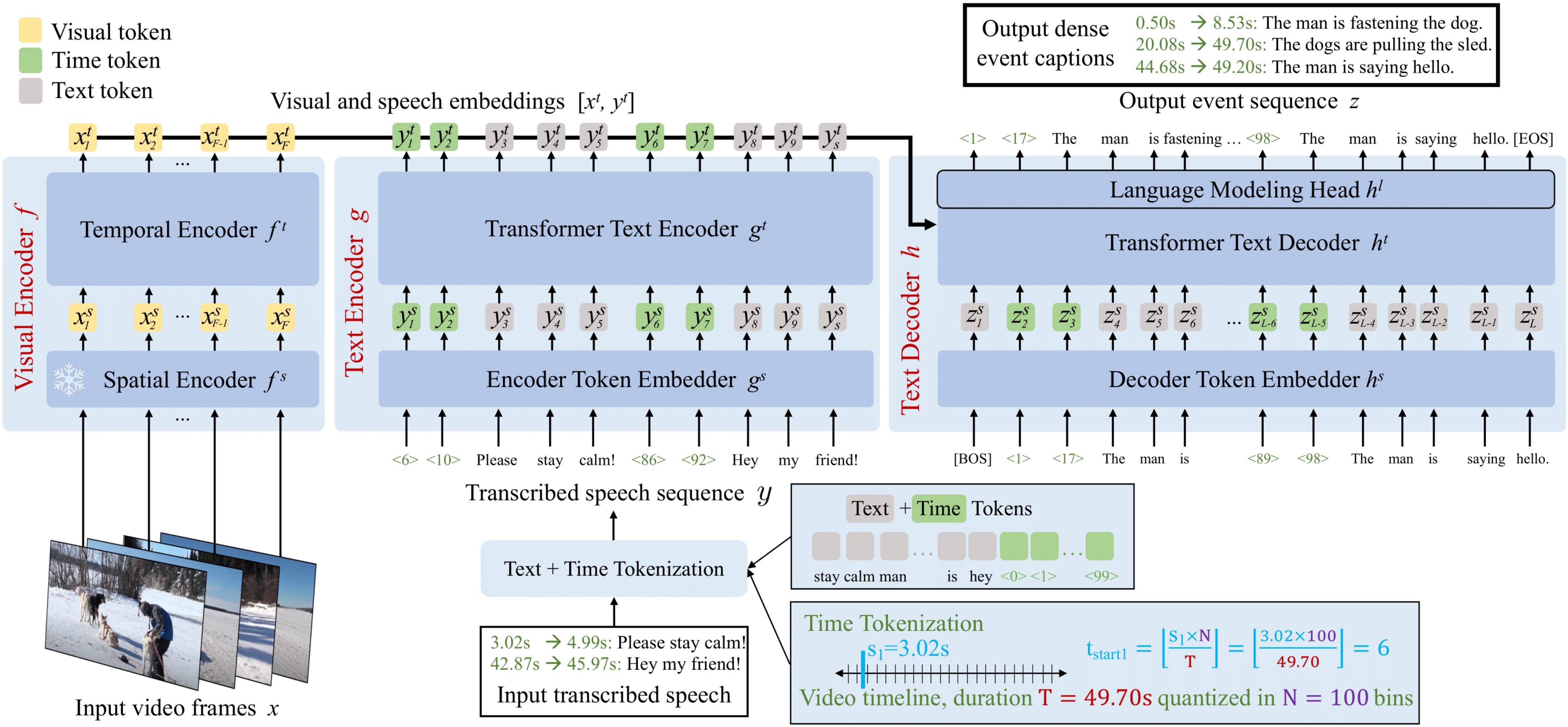
Goal: Use unlabeled narrated videos to train dense video captioning model




https://antoyang.github.io/vid2seq.html

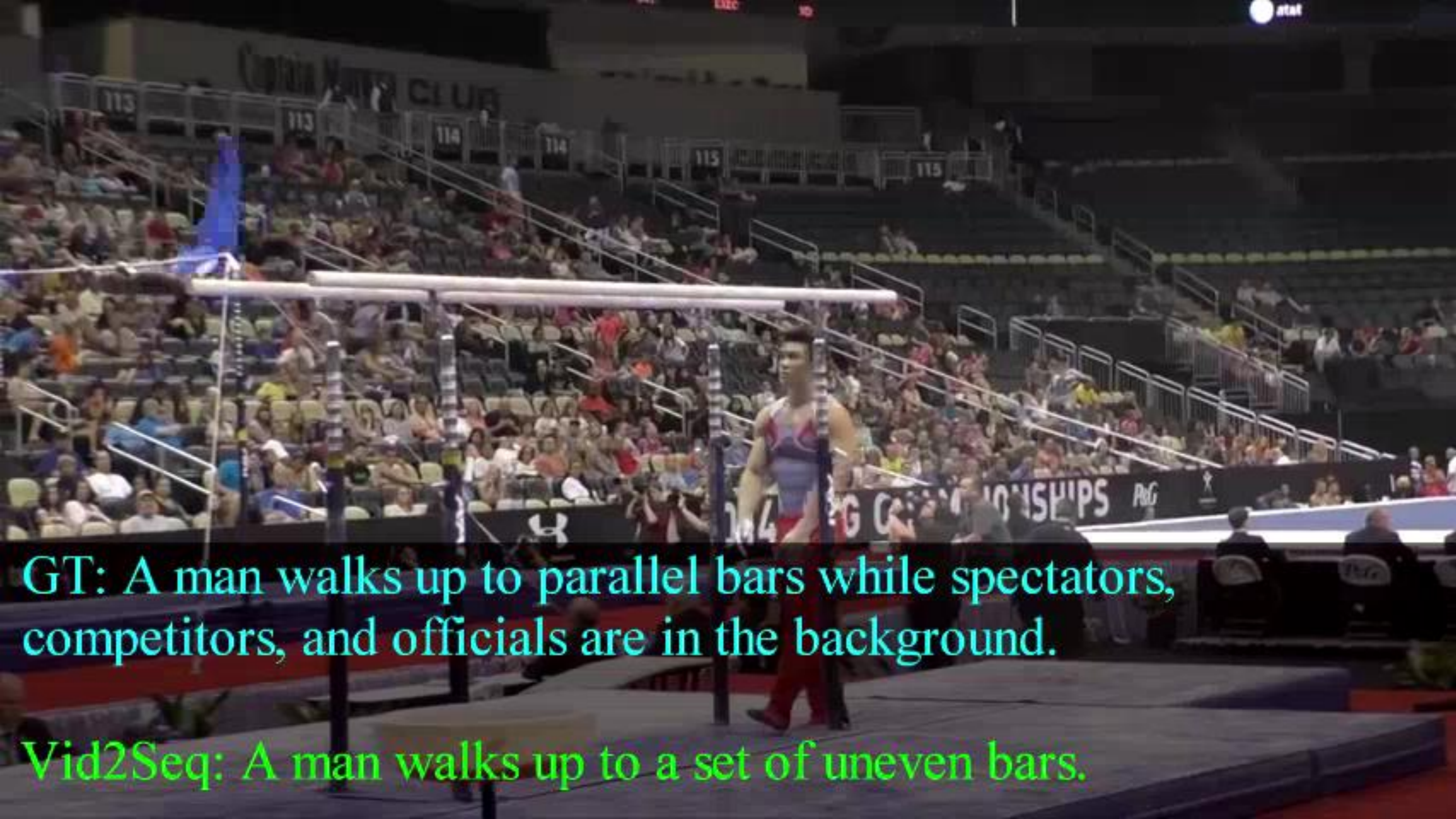
[Yang, Nagrani, Seo, Miech, Pont-Tuset, Laptev, Sivic and Schmid, CVPR 2023]

# Vid2Seq model




https://antoyang.github.io/vid2seq.html

[Yang, Nagrani, Seo, Miech, Pont-Tuset, Laptev, Sivic and Schmid, CVPR 2023]


GT: A man walks up to parallel bars while spectators, competitors, and officials are in the background.
Vid2Seq: A man walks up to a set of uneven bars.

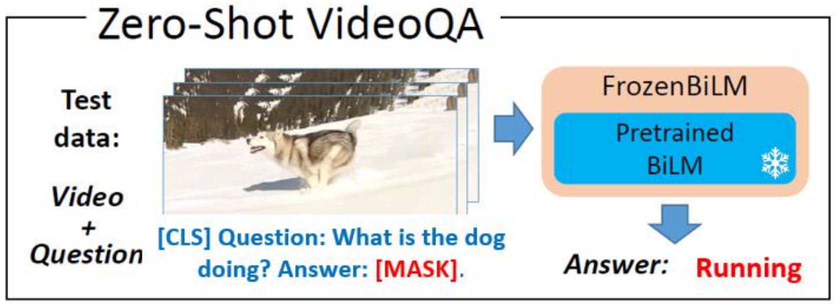

Vid2Seq: Trim off the excess fat of chicken breast
and cut it into halves.

# Is video understanding getting solved?

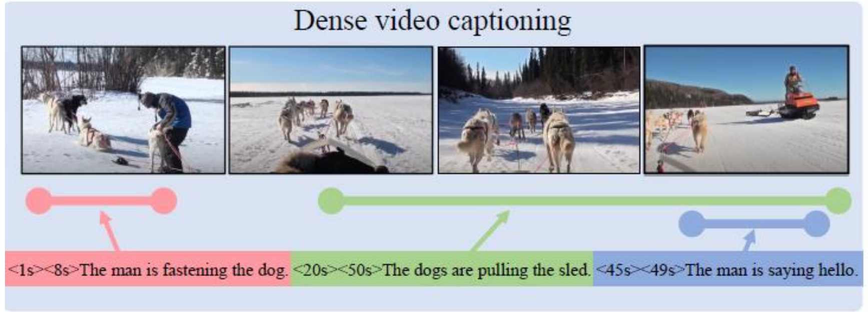
Park et al., CVPR19

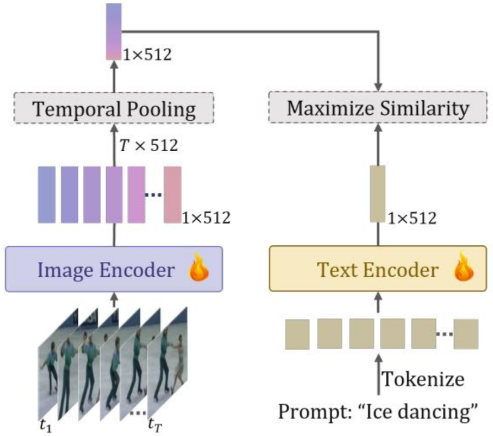
A small group of people are seen riding around in bumper cars and bumping into one another. The girl continues riding around the bumper car while others watch on the side. The girl finishes and walks away.

Yang et al., NeurIPS 2022




Yang et al., CVPR 2023




ViFi-CLIP Rasheed et al., 2023

With large-scale data and unsupervised training modern methods are getting excellent at associating video with language.

**Is this sufficient?**

# Open challenges in vision

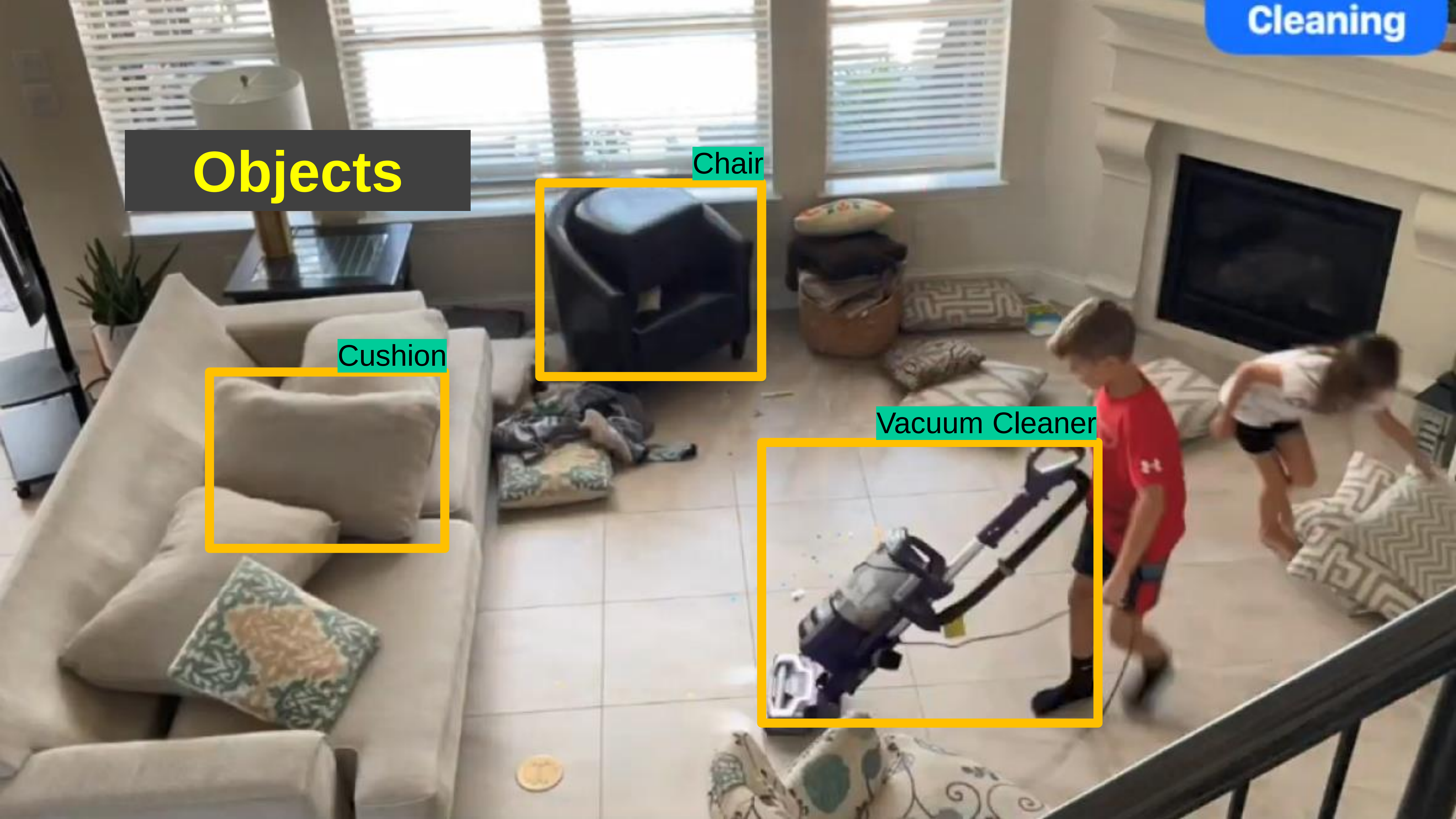
What are effects of certain actions on a given scene?

What happens if…?

…shaking an apple tree

…pulling tablecloth

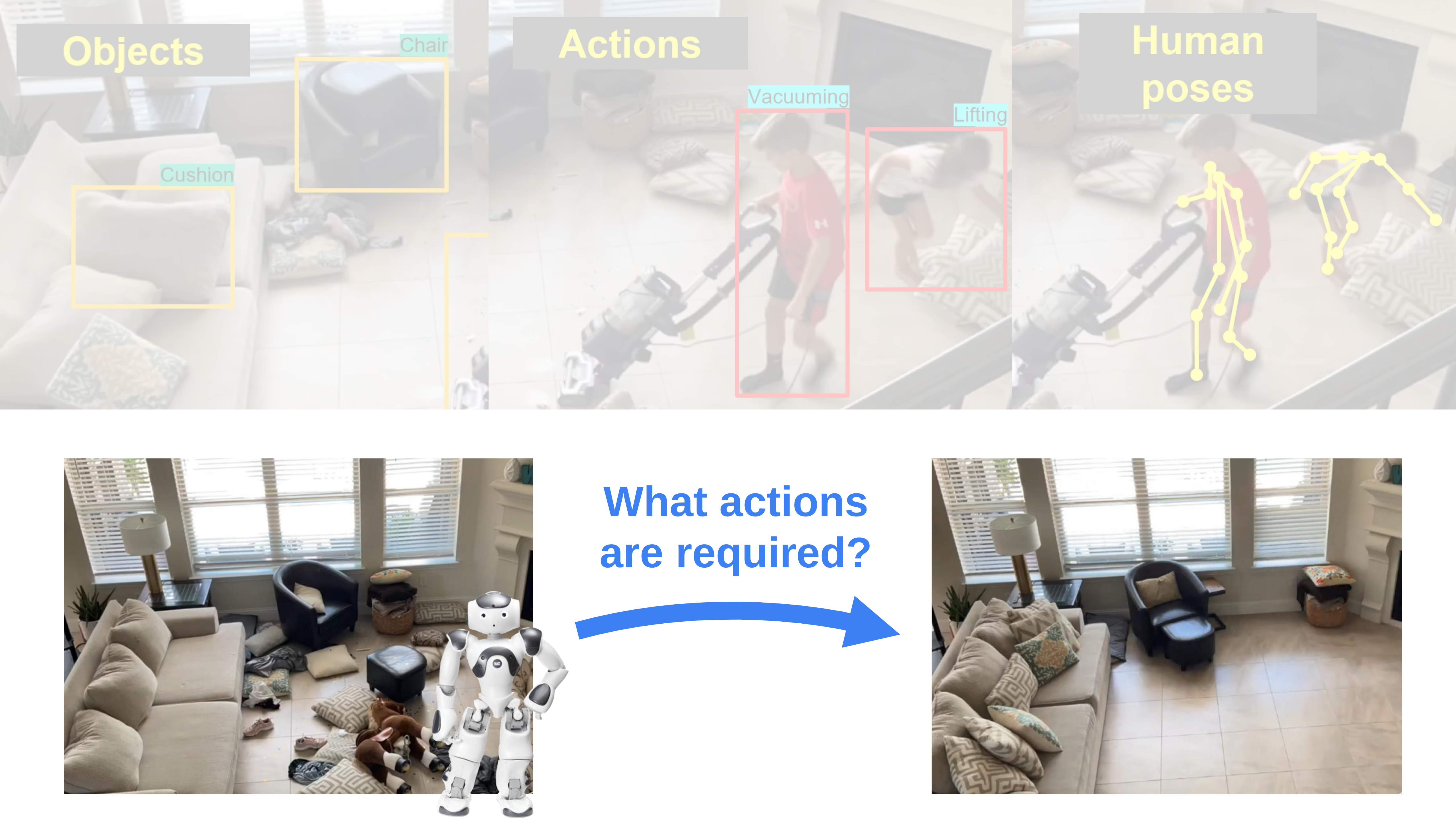

NASA


Cleaning
Objects
Chair
Cushion
Vacuum Cleaner

Cleaning
Actions
Vacuuming
Lifting

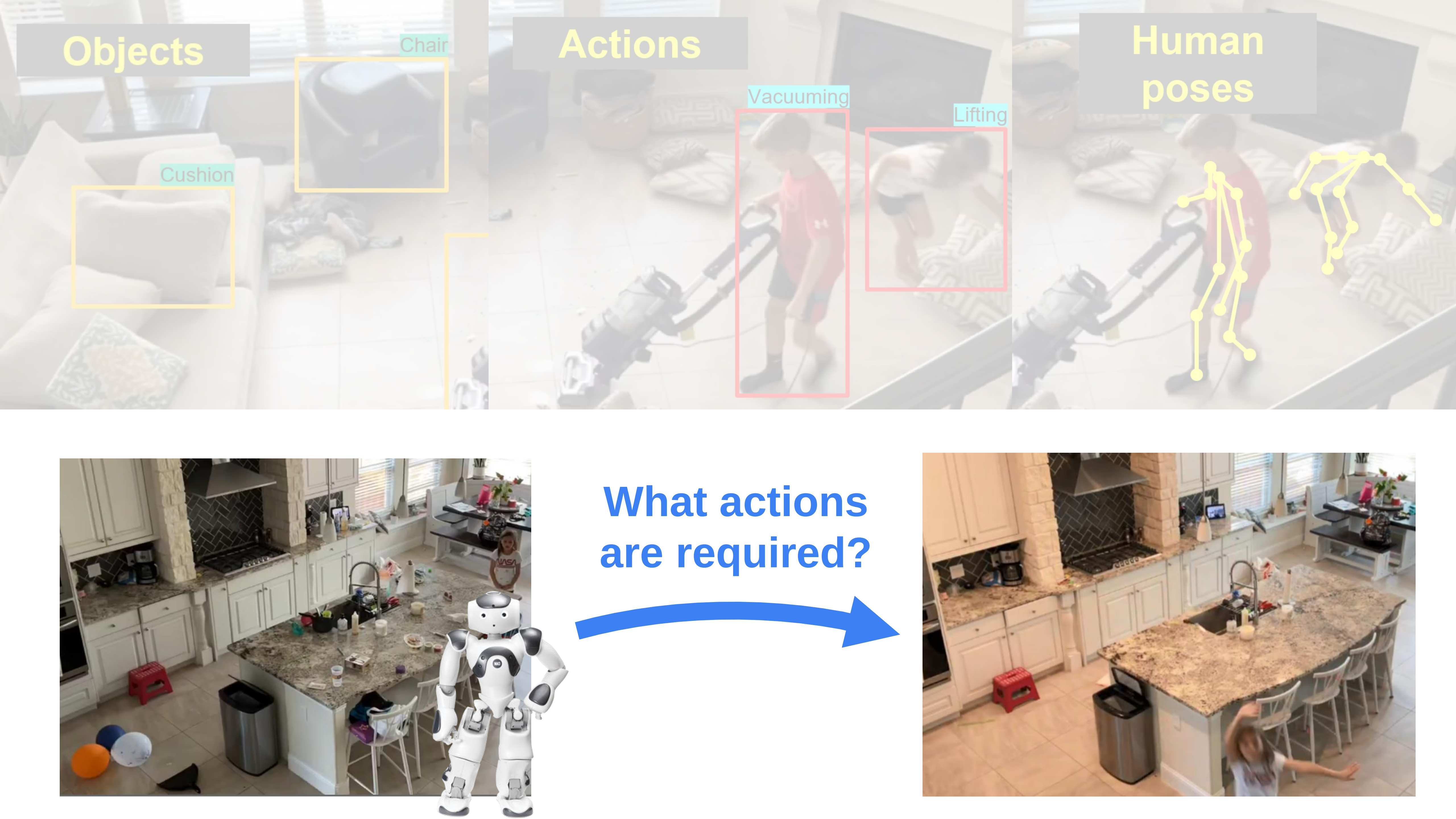
Cleaning
Human poses


Objects
Chair
Cushion
Actions
Vacuuming
Lifting
Human poses


Objects
Chair
Cushion
Actions
Vacuuming
Lifting
Human poses
What actions are required?

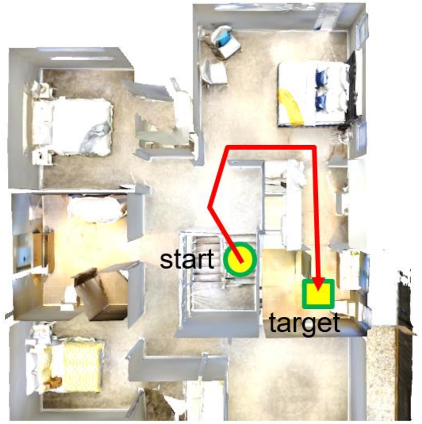
What actions are required?

What actions are required?


Objects
Chair
Cushion
Actions
Vacuuming
Lifting
Human
poses
What actions
are required?


Navigation
Manipulation

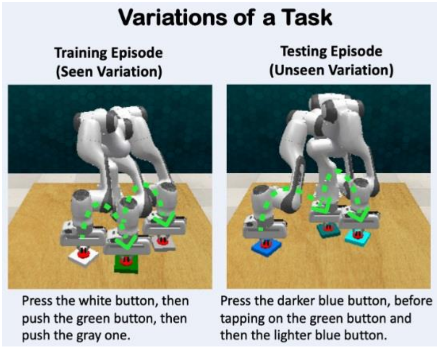
# Summary




**History Aware Multimodal Transformer for Vision-and-Language Navigation**, S. Chen, P.-L. Guhur, C. Schmid and I. Laptev; *in Proc. NeurIPS 2021*

**Object Goal Navigation with Recursive Implicit Maps**, S. Chen, T. Chabal, I. Laptev and C. Schmid; *In submission 2023*

Vision and language **navigation**

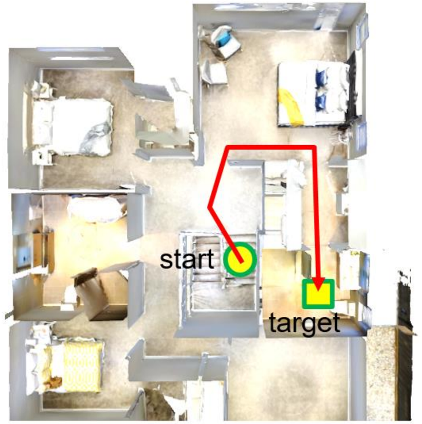



**Instruction-driven history-aware policies for robotic manipulations**, P.-L. Guhur, S. Chen, R. Garcia, M. Tapaswi, I. Laptev and C. Schmid; *in Proc. CoRL 2022*

**Robust visual sim-to-real transfer for robotic manipulation**, R. Garcia, R. Strudel, S. Chen, E. Arlaud, I. Laptev and C. Schmid. *In submission 2023*

Vision and language **manipulation**

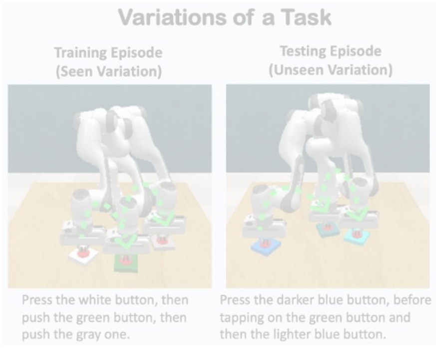
# Summary




**History Aware Multimodal Transformer for Vision-and-Language Navigation**, S. Chen, P.-L. Guhur, C. Schmid and I. Laptev; *in Proc. NeurIPS 2021*

**Object Goal Navigation with Recursive Implicit Maps**, S. Chen, T. Chabal, I. Laptev and C. Schmid; *In submission 2023*

Vision and language **navigation**




**Instruction-driven history-aware policies for robotic manipulations**, P.-L. Guhur, S. Chen, R. Garcia, M. Tapaswi, I. Laptev and C. Schmid; *in Proc. CoRL 2022*

**Robust visual sim-to-real transfer for robotic manipulation**, R. Garcia, R. Strudel, S. Chen, E. Arlaud, I. Laptev and C. Schmid. *In submission 2023*

Vision and language **manipulation**

# History Aware Multimodal Transformer for Vision-and-Language Navigation




**Shizhe Chen** Pierre-Louis Guhur Cordelia Schmid Ivan Laptev

NeurIPS 2021

**Webpage**:
https://cshizhe.github.io/projects/vln_hamt.html

# VLN Challenges: Modeling history

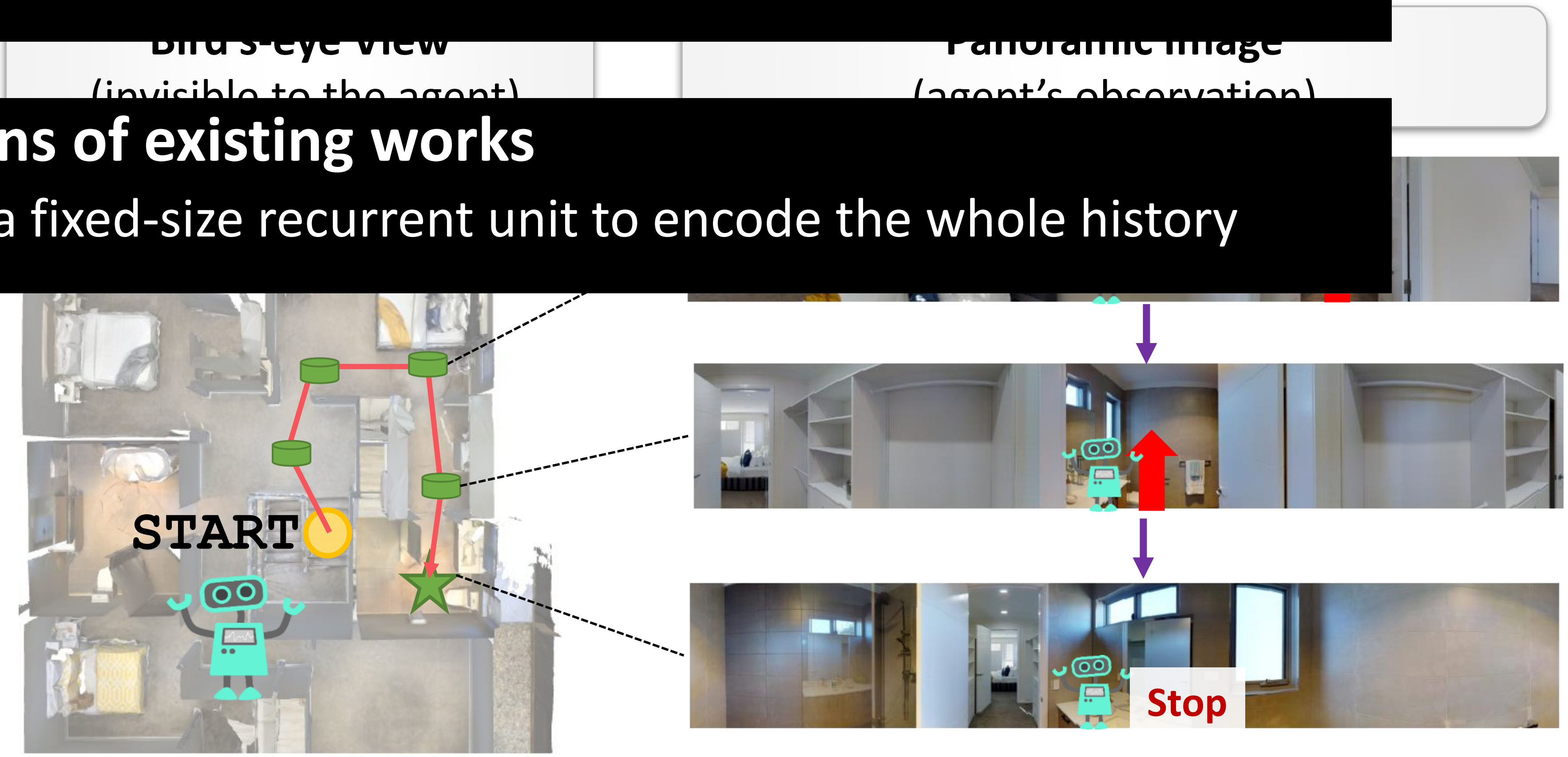
**Keeping track of the navigation state**

Environment understanding

Instruction grounding

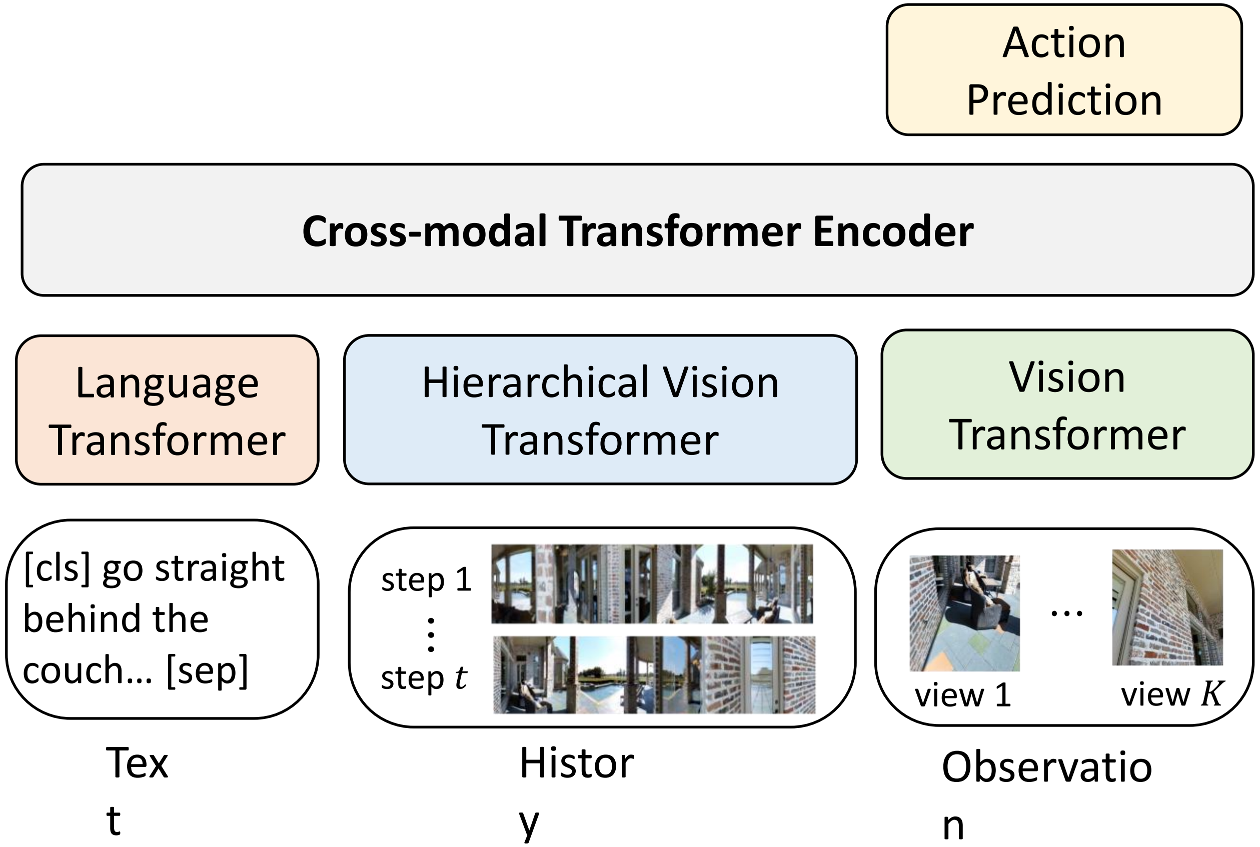
- **Limitations of existing works**
  - Adopt a fixed-size recurrent unit to encode the whole history

Turn left and continue
up the st
Go straig
the bedr
the right
past the bed.
Turn right again and
go through the closet.
Continue straight, into
the bathroom.
Wait right there, in
front of the mirror.



# Our Proposed Model: HAMT

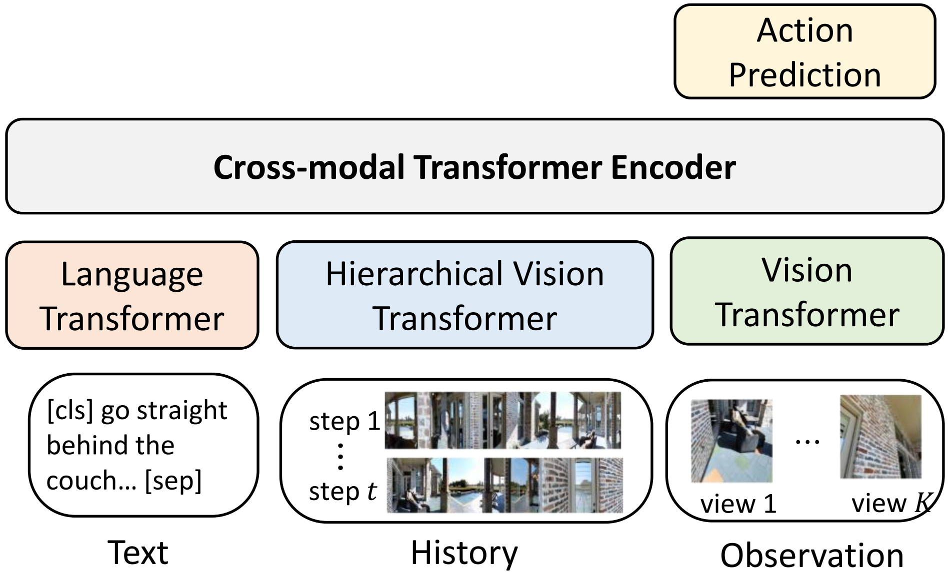
History Aware Multimodal Transformer (HAMT)




A fully transformer-based architecture for multimodal decision making

# Our Proposed Model: HAMT

Long-horizon history modelling
Learn dependency of all panoramic observations and actions in history sequence
End-to-end optimization for visual representation
Fully transformer-based architecture allows efficient training

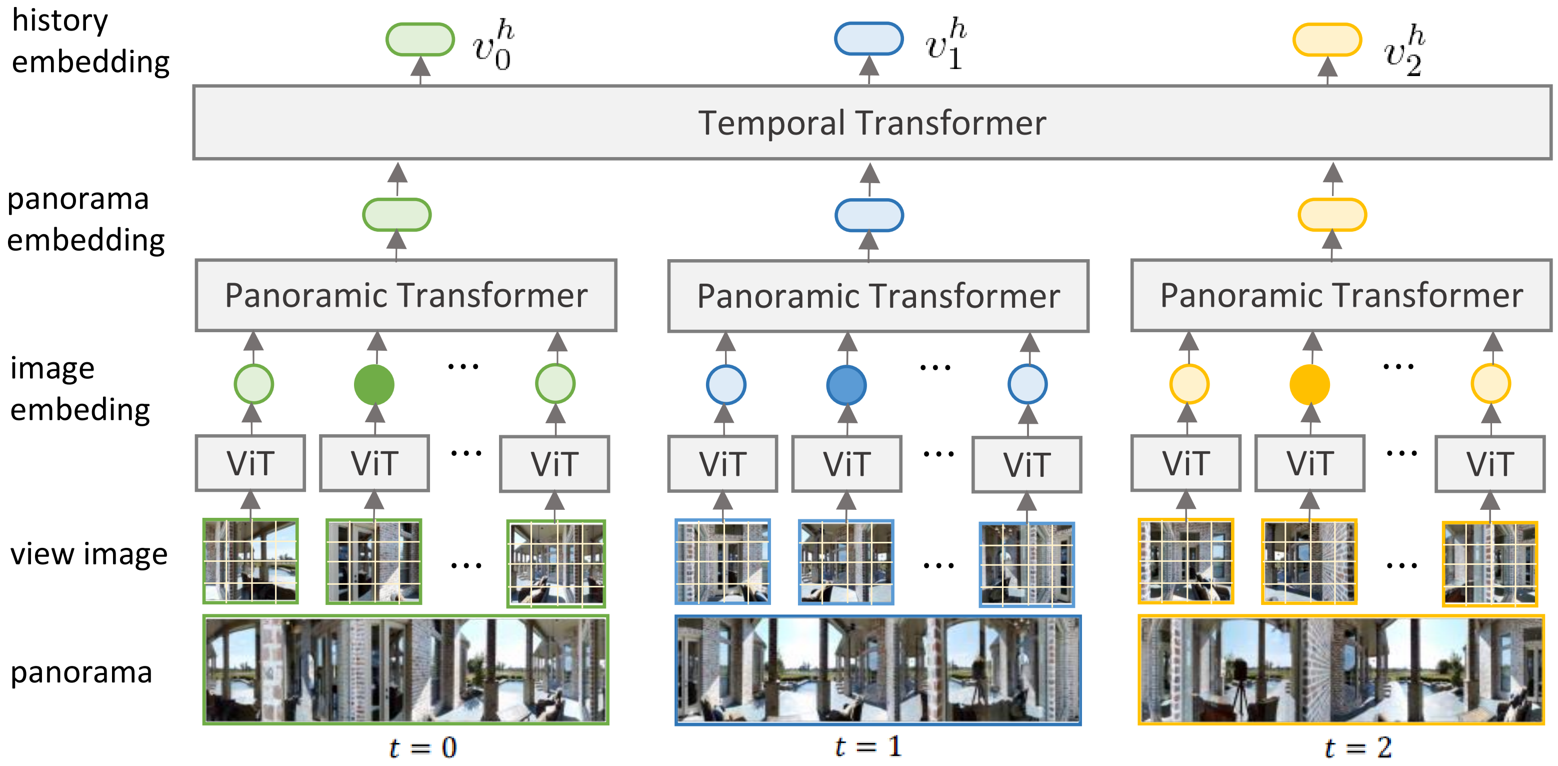



**PROBLEMS**

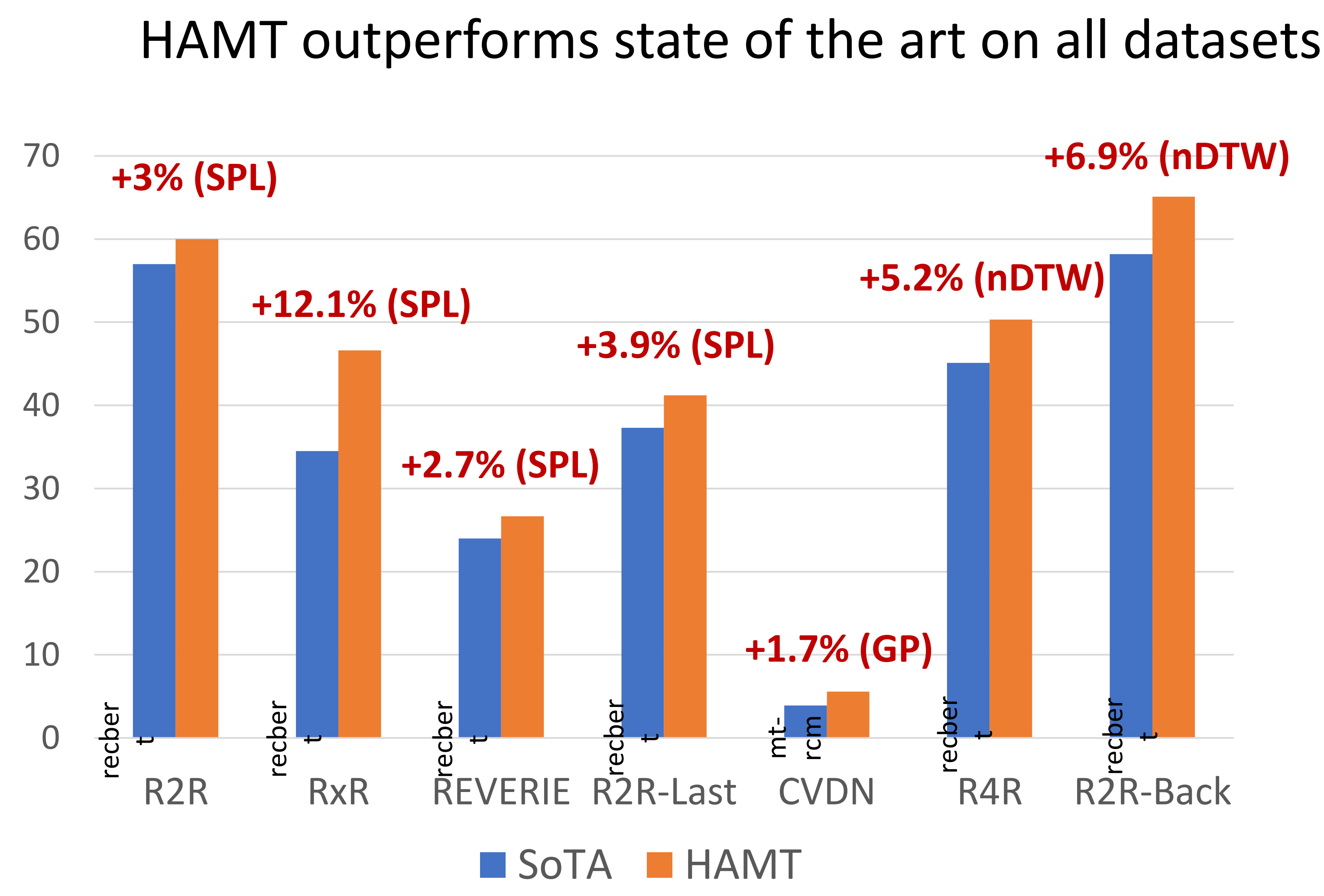
- Computationally expensive to encode all panoramas
  - K views, T steps → $O(K^2T^2)$
- The action prediction task alone might be insufficient to learn generalizable models

# HAMT: Hierarchical History Encoding

ViT for single view image encoding
Panoramic Transformer for spatial relation encoding within panorama
Temporal Transformer for temporal relation encoding across panoramas



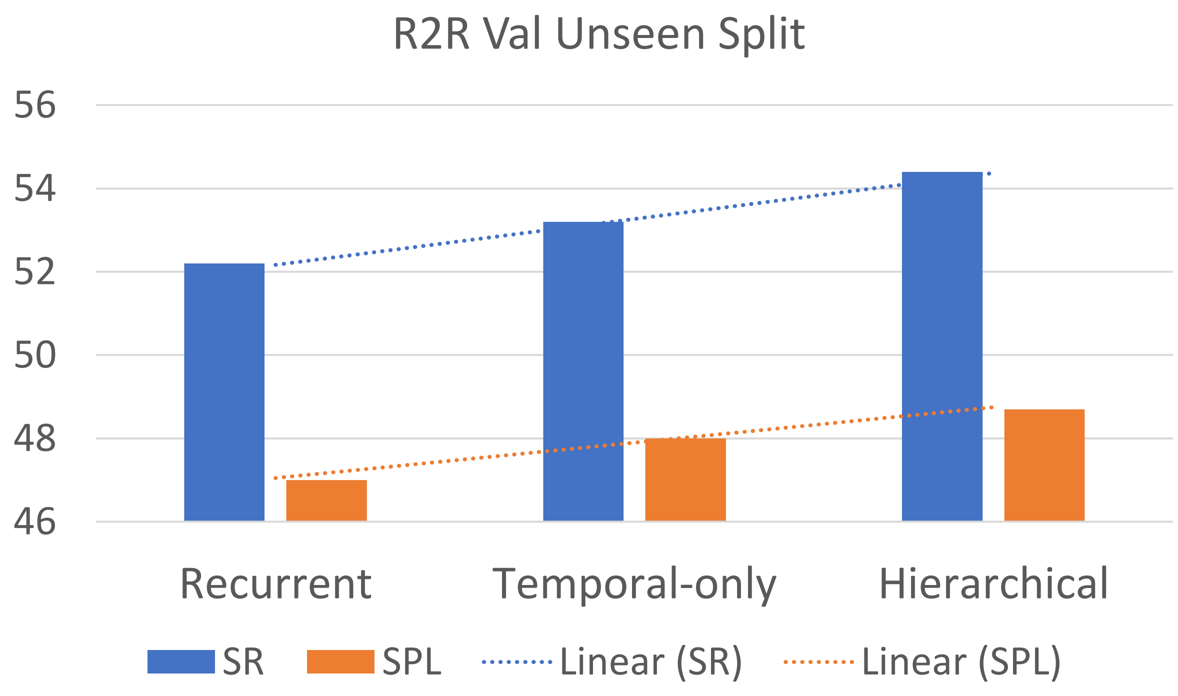
# Experiments: Comparison with SoTA



# Experiments: Ablation

How important is the history encoding?

- Recurrent: a fixed-size vector to encode the whole history
- Temporal-only: select only one view per panorama to improve efficiency
- **Hierarchical: hierarchically encode all panoramas**



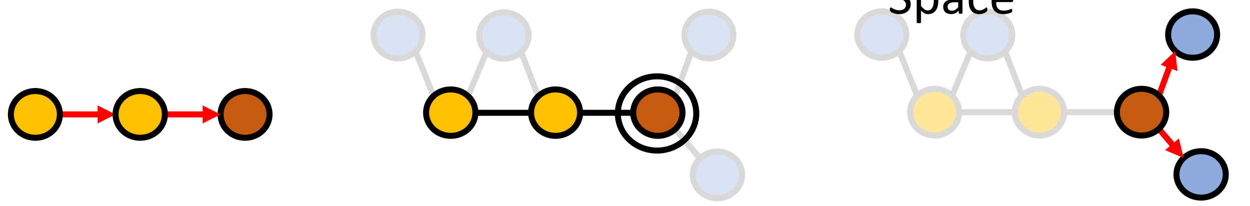
# Limitations of HAMT

Navigation Memory

Visual Memory

Action Space

HAMT



**Sequence**
no structure of the house

**Local actions**
hard to backtrack many steps

Current location

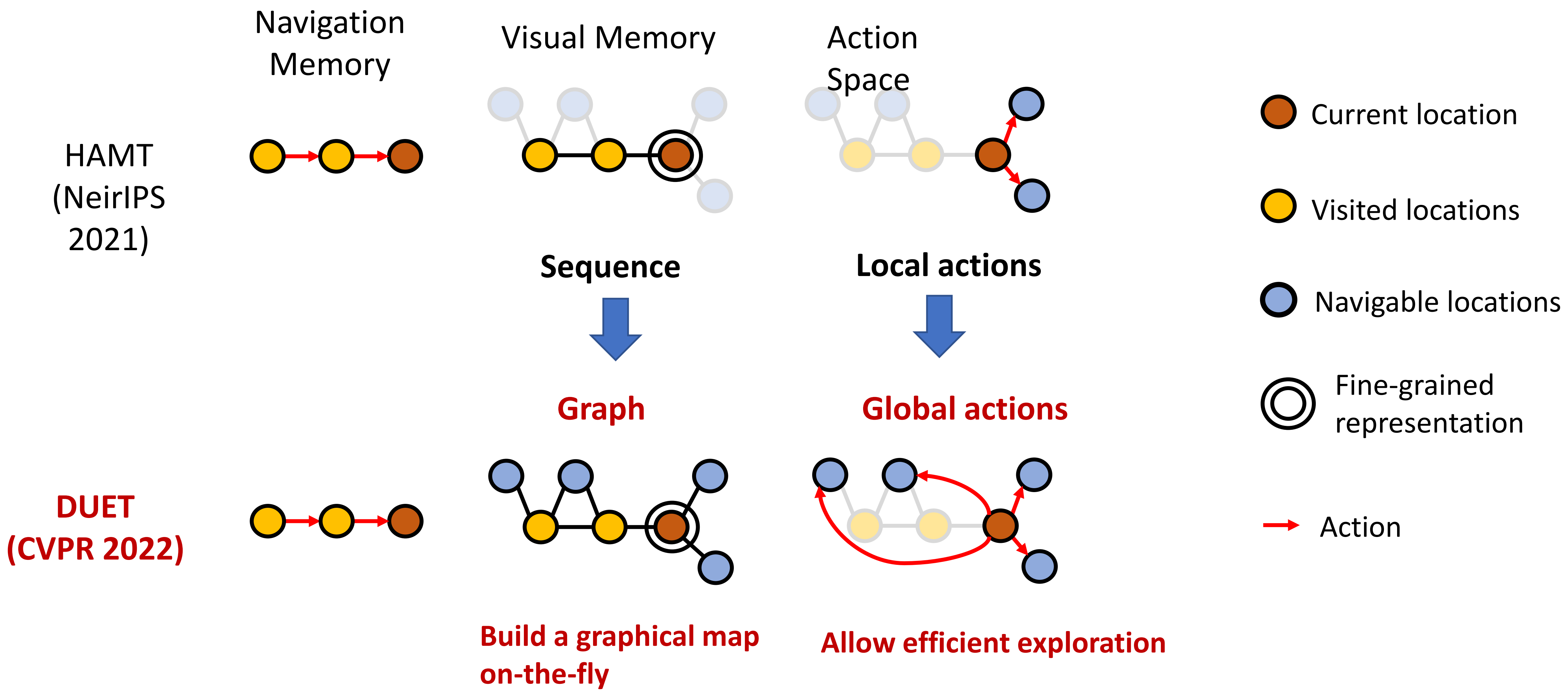
Visited locations

Navigable locations

Fine-grained representation

Action

# Improving HAMT with Structured Memory



# DUET: Experimental Results

REVERIE dataset

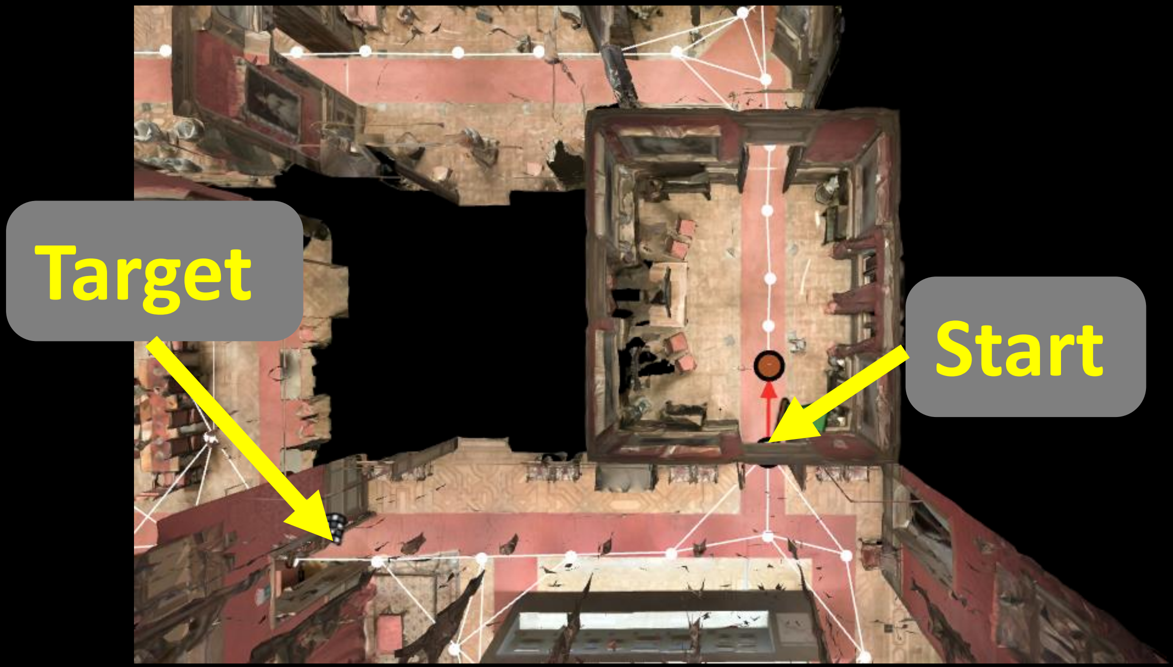
| | SR | SPL | RGS | RGSPL |
|---|---|---|---|---|
| HAMT | 30.40 | 26.67 | 14.88 | 13.08 |
| DUET | **52.51** | **36.06** | **31.88** | **22.06** |

- 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** |

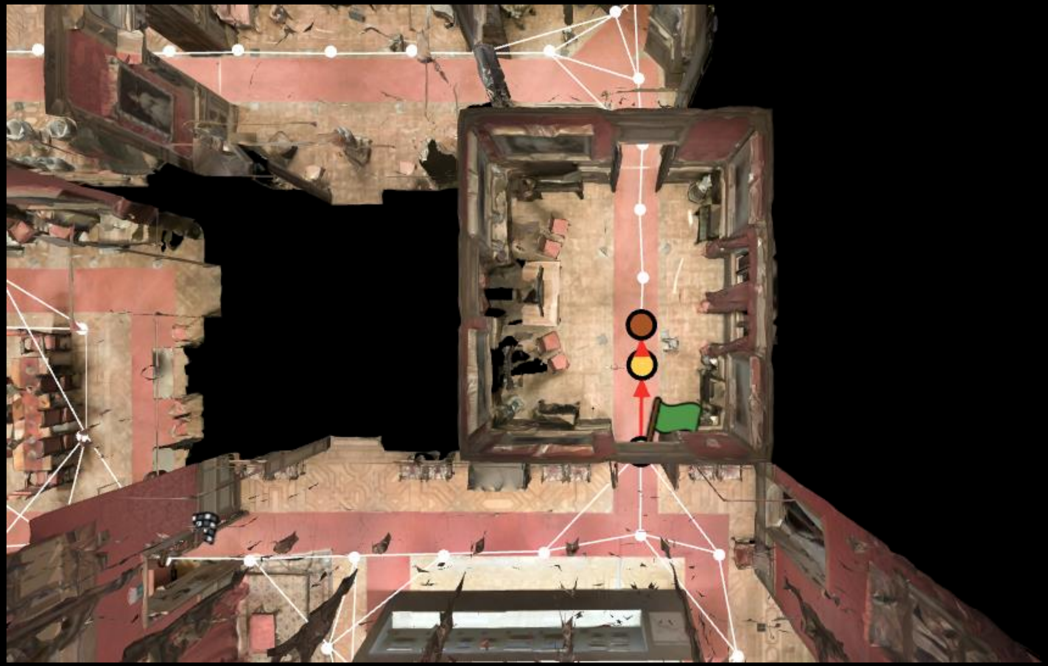
- **Winner of VLN Challenges** hosted in Human Interaction for Robotics Navigation Workshop at ICCV 2021

Instruction: Exit the roped off hall, follow the red carpet, turn right, continue straight down the red carpet, enter room at the end, stop once inside the room.

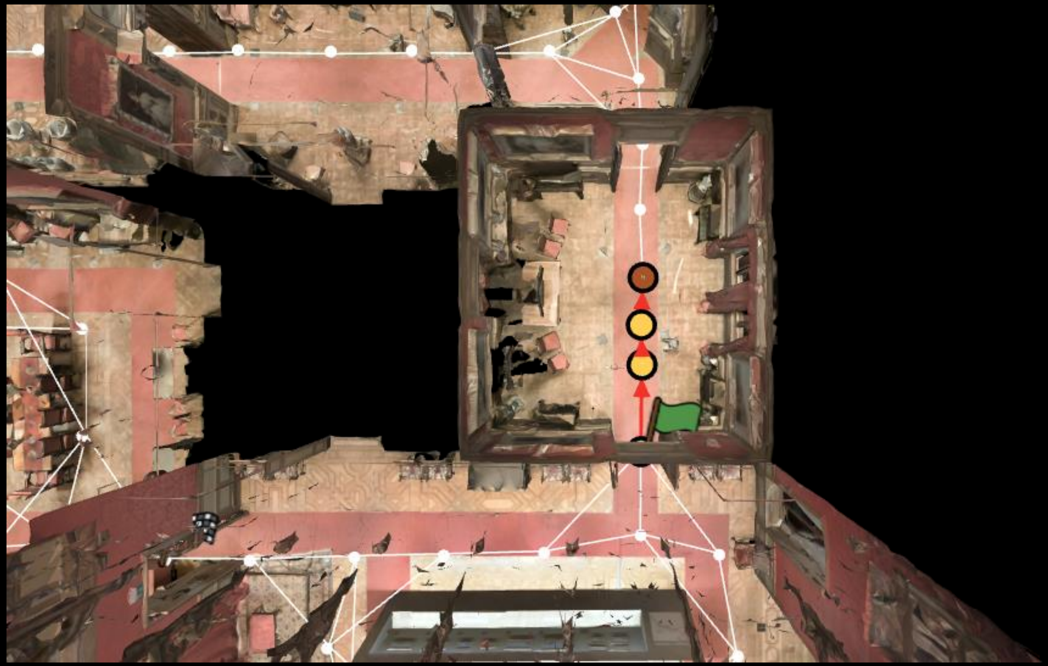


Instruction: **Exit the roped off hall, follow the red carpet**, turn right, continue straight down the red carpet, enter room at the end, stop once inside the room.

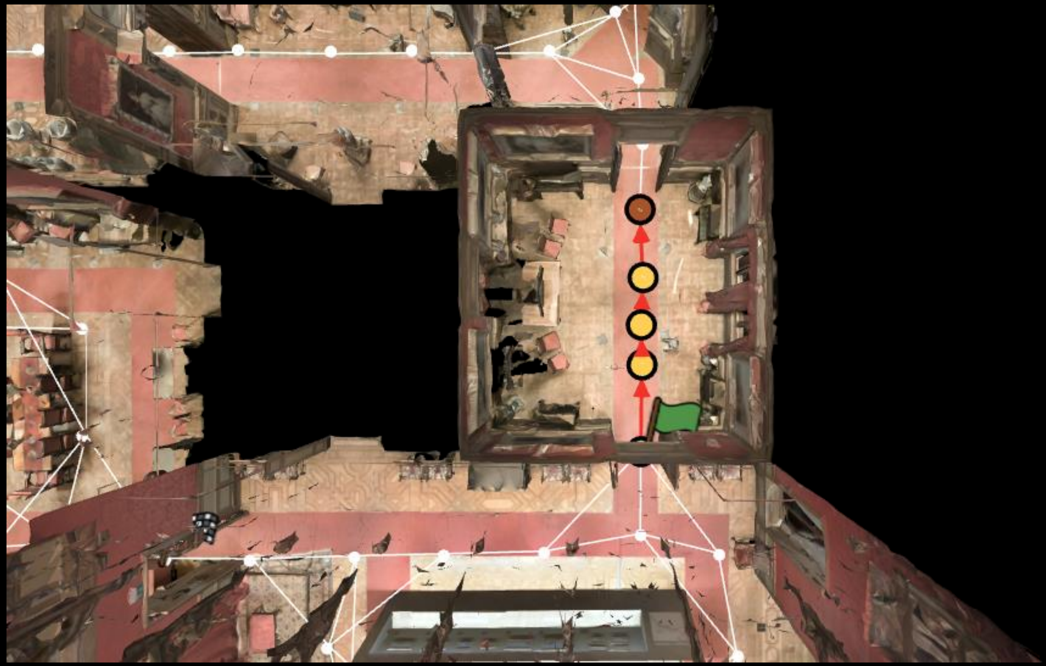
Instruction: **Exit the roped off hall, follow the red carpet**, turn right, continue straight down the red carpet, enter room at the end, stop once inside the room.



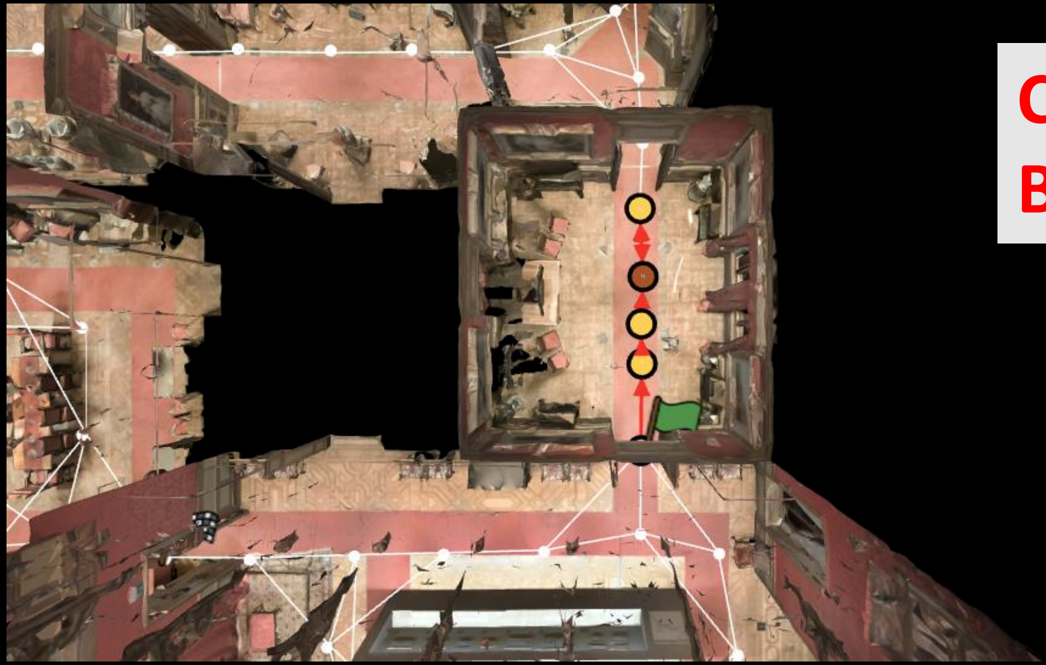
Instruction: **Exit the roped off hall, follow the red carpet**, turn right, continue straight down the red carpet, enter room at the end, stop once inside the room.



Instruction: **Exit the roped off hall, follow the red carpet**, turn right, continue straight down the red carpet, enter room at the end, stop once inside the room.

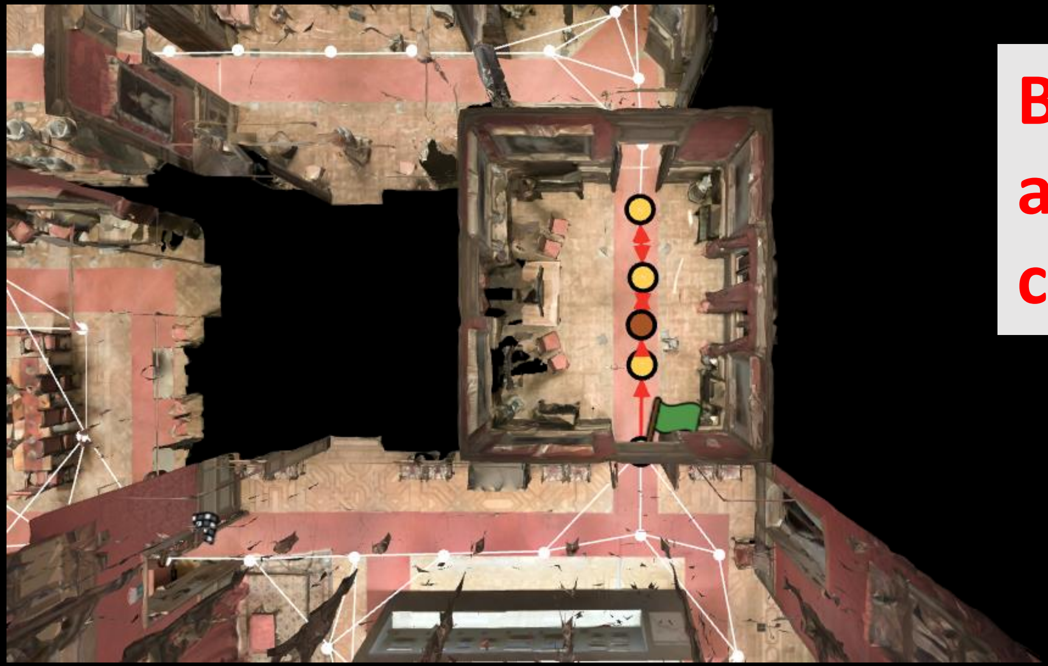


Instruction: Exit the roped off hall, follow the red carpet, **turn right**, continue straight down the red carpet, enter room at the end, stop once inside the room.



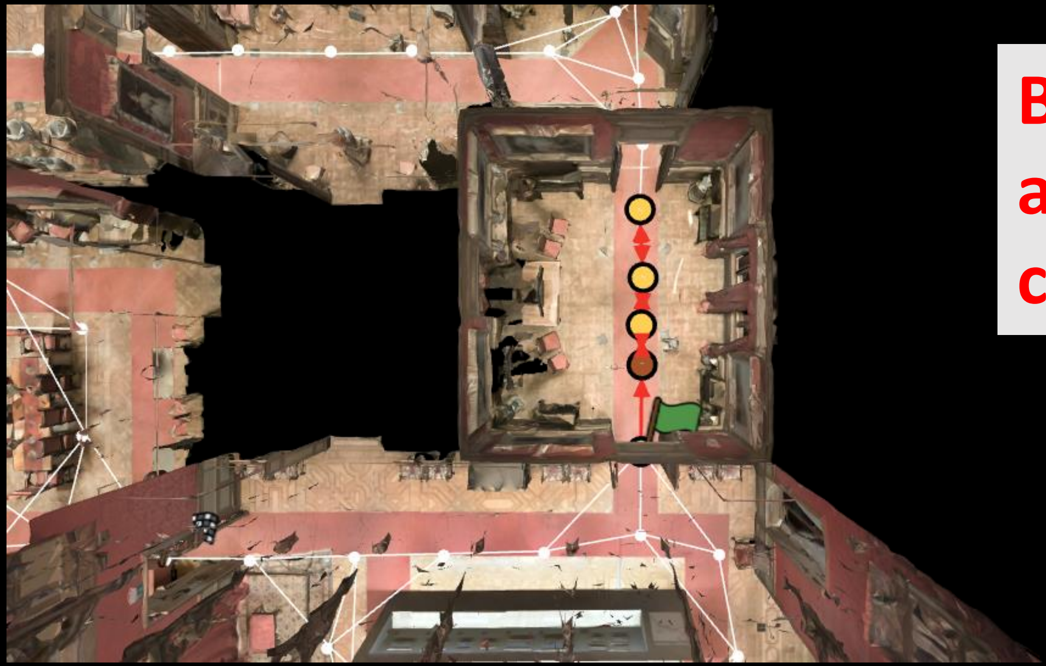
**Cannot turn right.**
**Back Track**

Instruction: Exit the roped off hall, follow the red carpet, turn right, continue straight down the red carpet, enter room at the end, stop once inside the room.



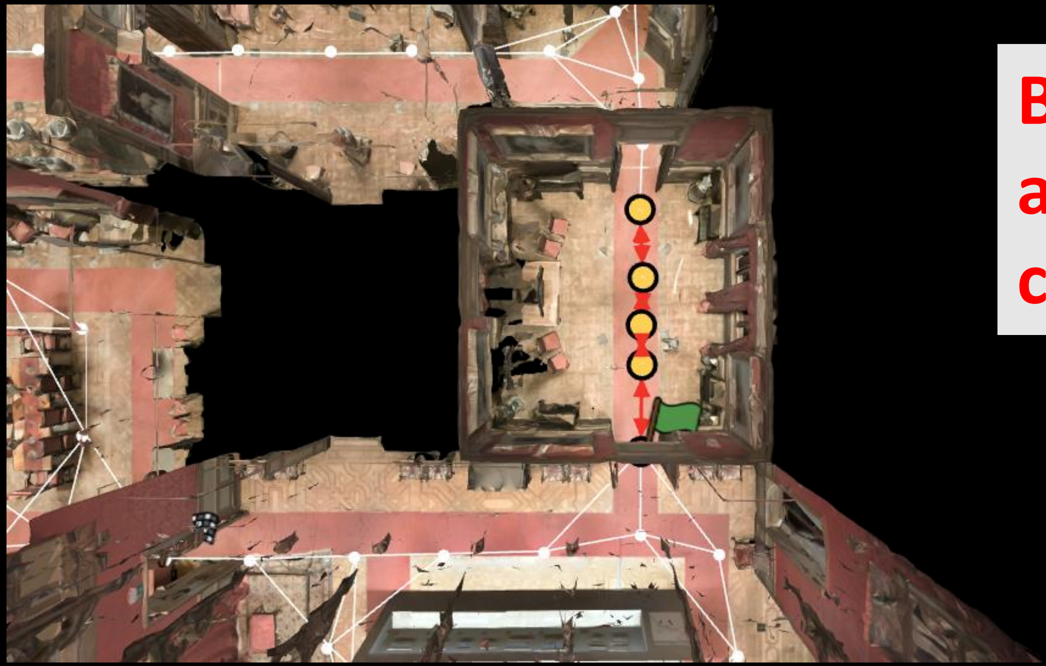
**Back tracking according to the constructed map.**

Instruction: Exit the roped off hall, follow the red carpet, turn right, continue straight down the red carpet, enter room at the end, stop once inside the room.



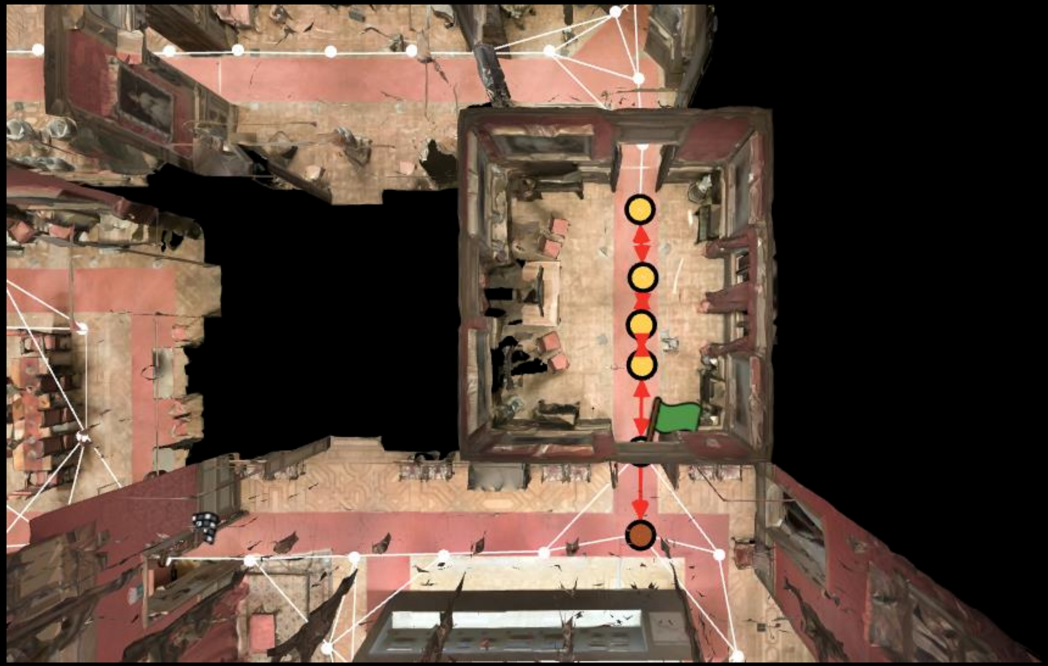
**Back tracking according to the constructed map.**

Instruction: Exit the roped off hall, follow the red carpet, turn right, continue straight down the red carpet, enter room at the end, stop once inside the room.

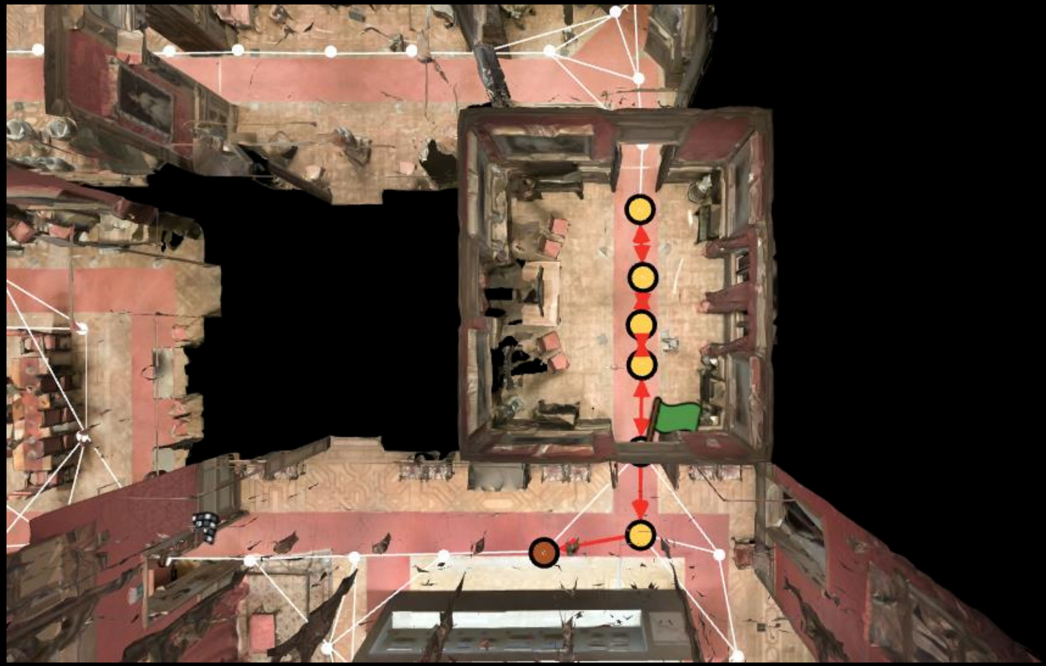


**Back tracking according to the constructed map.**

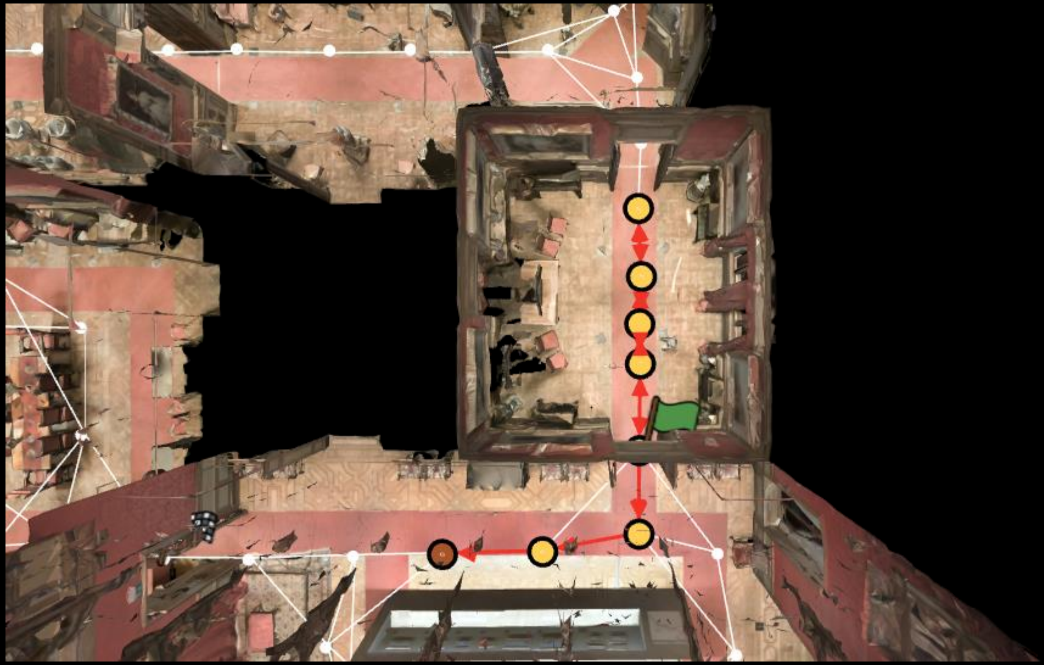
Instruction: **Exit the roped off hall, follow the red carpet**, turn right, continue straight down the red carpet, enter room at the end, stop once inside the room.



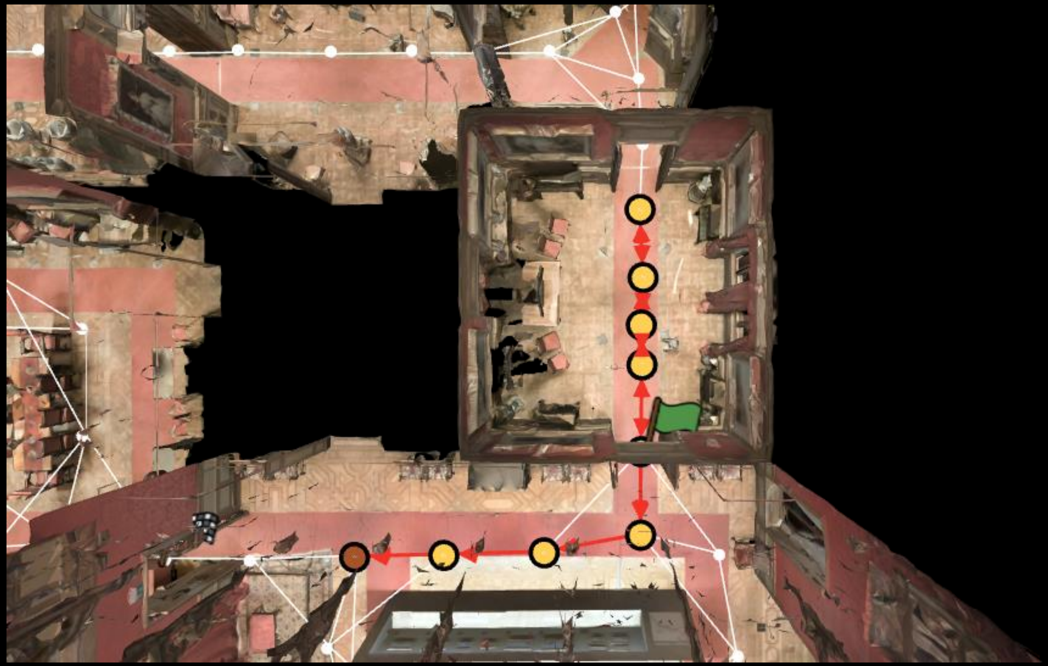
Instruction: Exit the roped off hall, follow the red carpet, **turn right**, continue straight down the red carpet, enter room at the end, stop once inside the room.



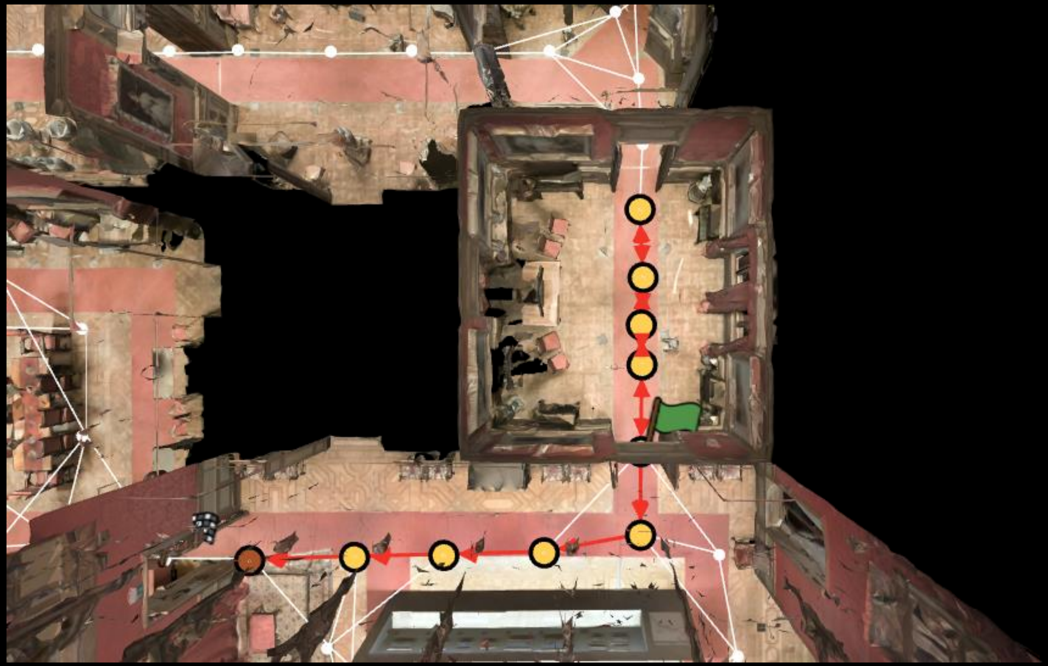
Instruction: Exit the roped off hall, follow the red carpet, turn right, **continue straight down the red carpet,** enter room at the end, stop once inside the room.



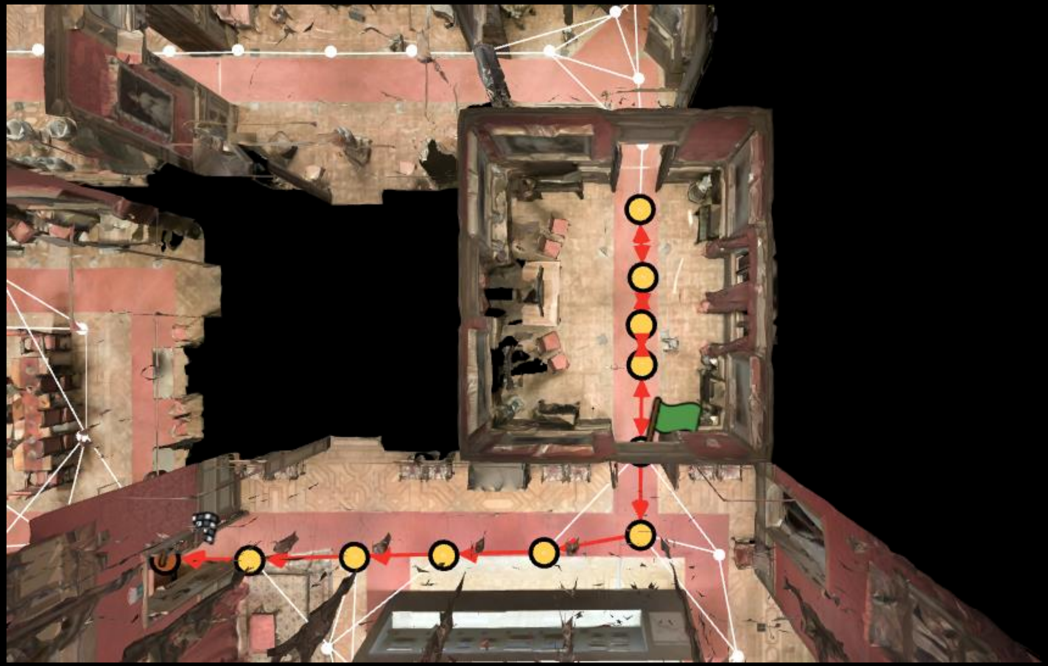
Instruction: Exit the roped off hall, follow the red carpet, turn right, **continue straight down the red carpet,** enter room at the end, stop once inside the room.



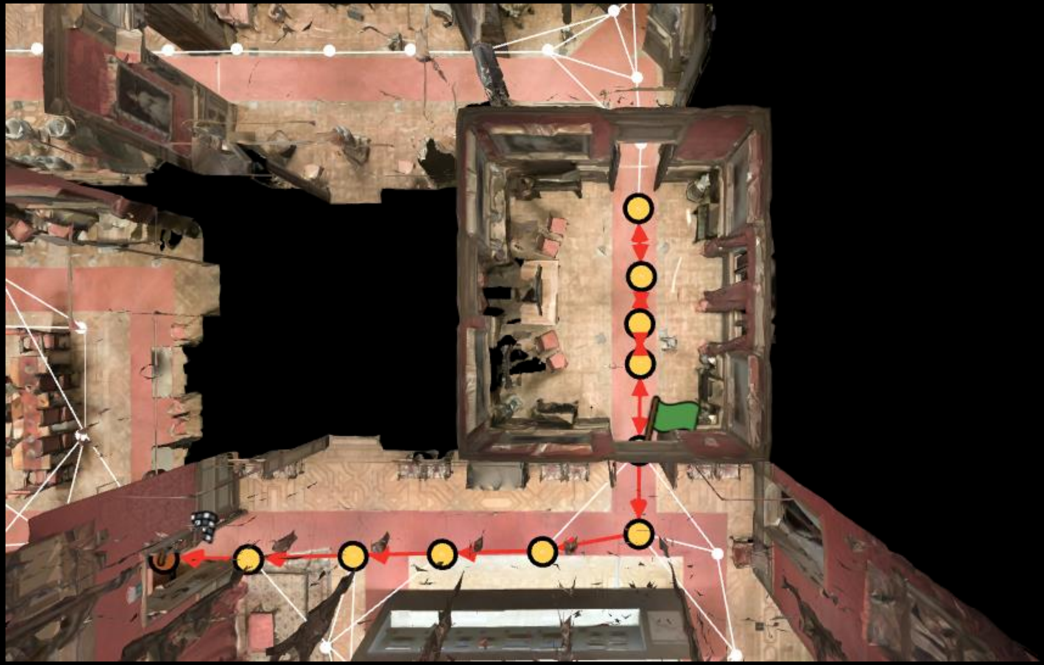
Instruction: Exit the roped off hall, follow the red carpet, turn right, **continue straight down the red carpet,** enter room at the end, stop once inside the room.



Instruction: Exit the roped off hall, follow the red carpet, turn right, continue straight down the red carpet, **enter room at the end, stop once inside the room.**



Instruction: Exit the roped off hall, follow the red carpet, turn right, continue straight down the red carpet, **enter room at the end, stop once inside the room.**

# Object Goal Navigation with Recursive Implicit Maps

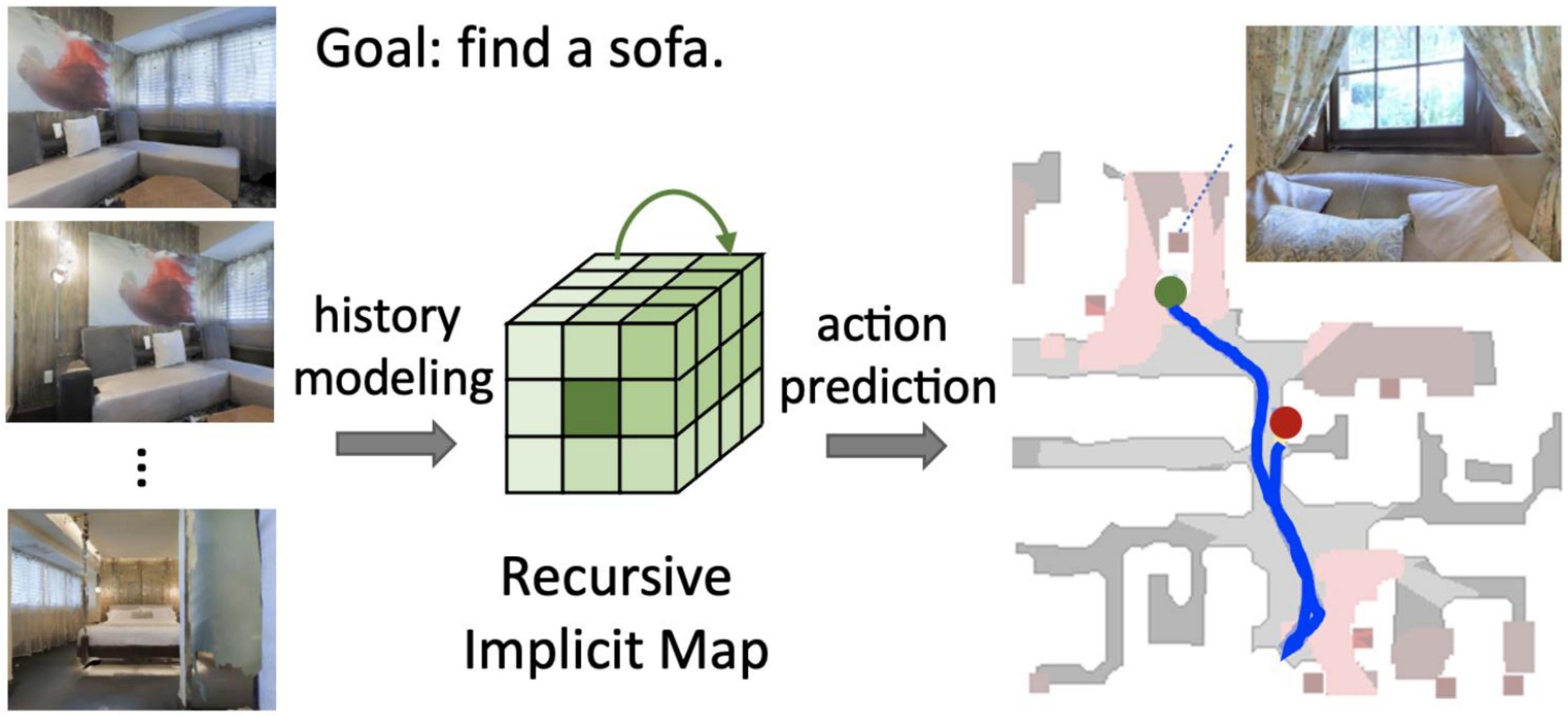




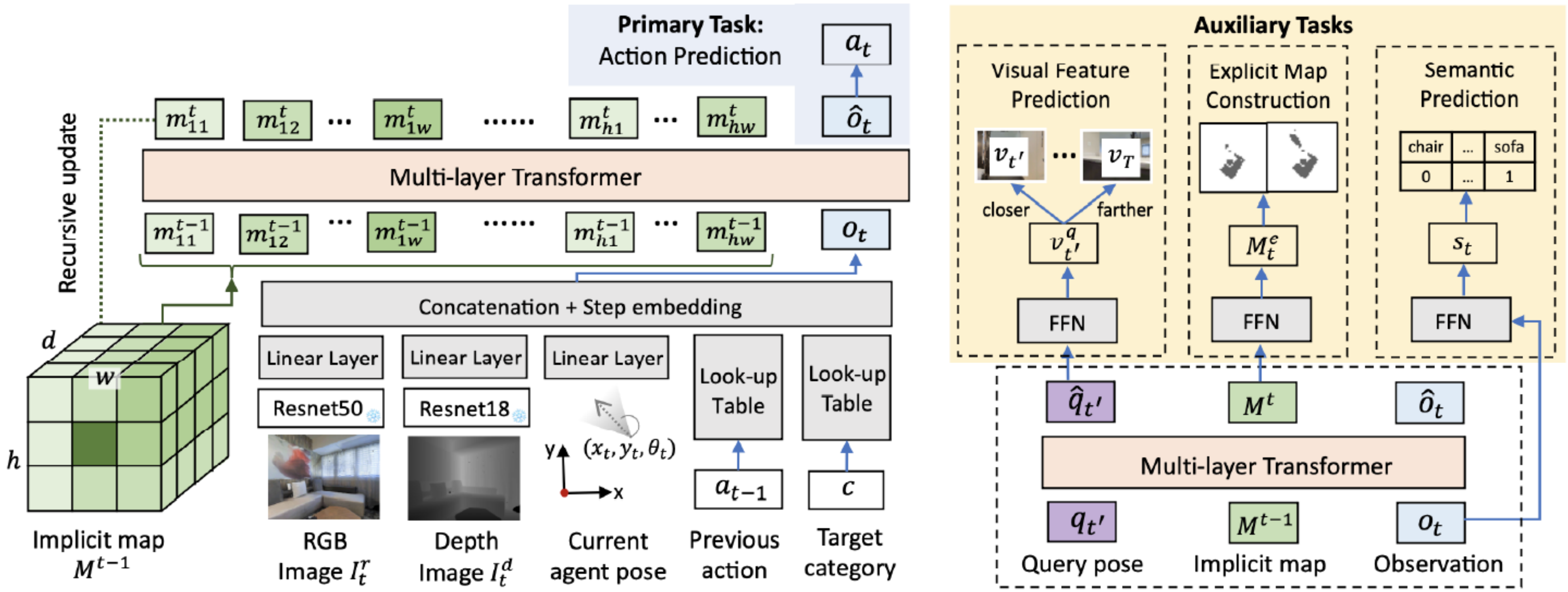
**Shizhe Chen** Thomas Chabal Cordelia Schmid Ivan Laptev


In submission 2023

# Object Navigation model with Recursive Implicit Map



# Object Navigation model with Recursive Implicit Map




| | Memory size | SR | SPL | SoftSPL |
|---|---|---|---|---|
| Recurrent state | $1 \times d$ | 38.95 | 11.09 | 16.35 |
| Episodic sequence | $T \times d$ | 44.51 | 14.17 | 19.35 |
| Recursive implicit map | $h \times w \times d$ | **47.74** | **15.12** | **20.51** |

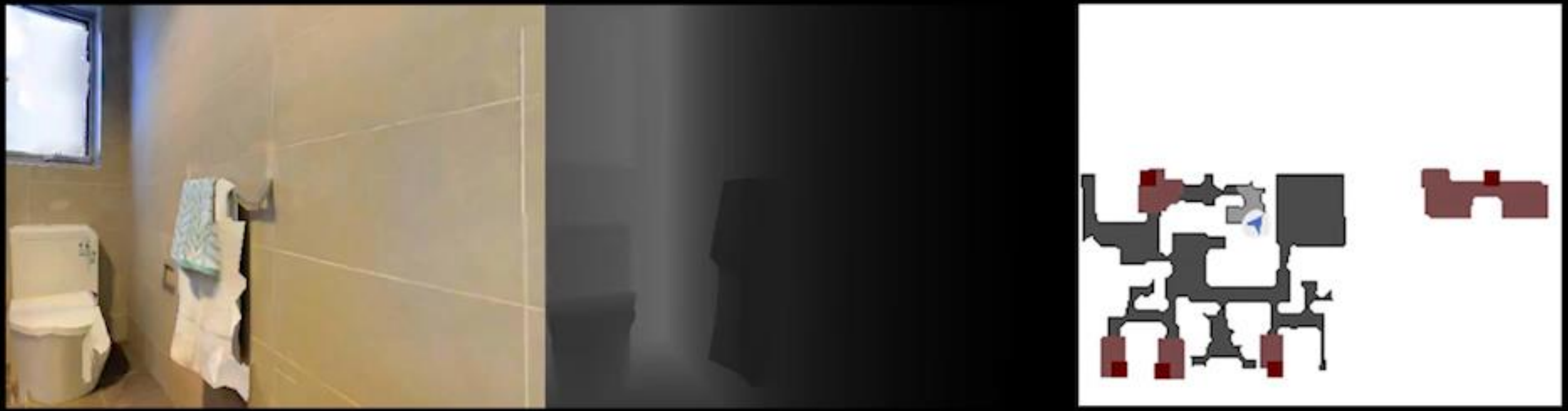
# Object Goal Navigation with Recursive Implicit Maps

Shizhe Chen, Thomas Chabal, Ivan Laptev and Cordelia Schmid

# Examples in simulation: successful cases

**Target:** *"cabinet"*

**Target:** *"chest of drawer"*



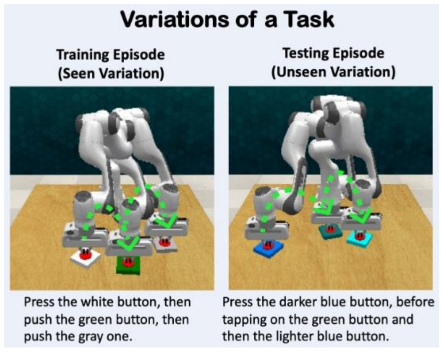
# Real world examples

# Summary


**Think Global, Act Local: Dual-scale Graph Transformer for Vision-and-Language Navigation**, S. Chen, P.-L. Guhur, M. Tapaswi, C. Schmid and I. Laptev; *in Proc. CVPR 2022*

**Object Goal Navigation with Recursive Implicit Maps**, S. Chen, T. Chabal, I. Laptev and C. Schmid; *In submission 2023*

Vision and language **navigation**




**Instruction-driven history-aware policies for robotic manipulations**, P.-L. Guhur, S. Chen, R. Garcia, M. Tapaswi, I. Laptev and C. Schmid; *in Proc. CoRL 2022*

**Robust visual sim-to-real transfer for robotic manipulation**, R. Garcia, R. Strudel, S. Chen, E. Arlaud, I. Laptev and C. Schmid. *In submission 2023*

Vision and language **manipulation**

# Instruction-driven History-aware Policies for Robotic Manipulation







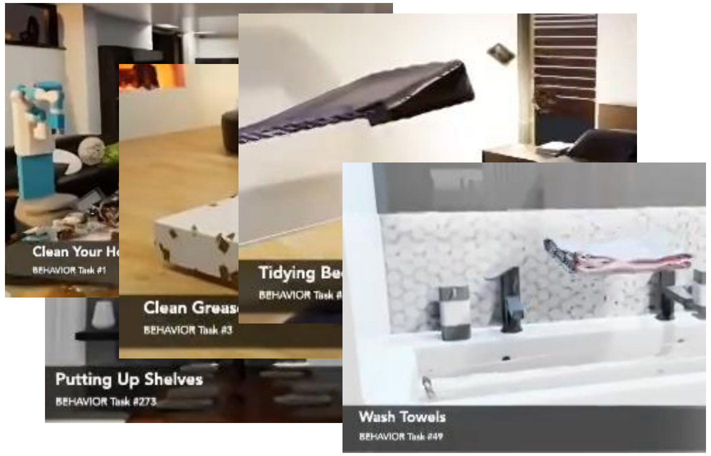
Pierre-Louis Guhur[1] Shizhe Chen[1] Ricardo Garcia Pinel[1] Makarand Tapaswi[1,2] Ivan Laptev[1] Cordelia Schmid[1]

**[1]Inria, École normale supérieure, CNRS, PSL Research University, Paris, France,**
**[2]IIIT Hyderabad, India**


Project page: [https://guhur.github.io/hiveformer/](https://guhur.github.io/hiveformer/)

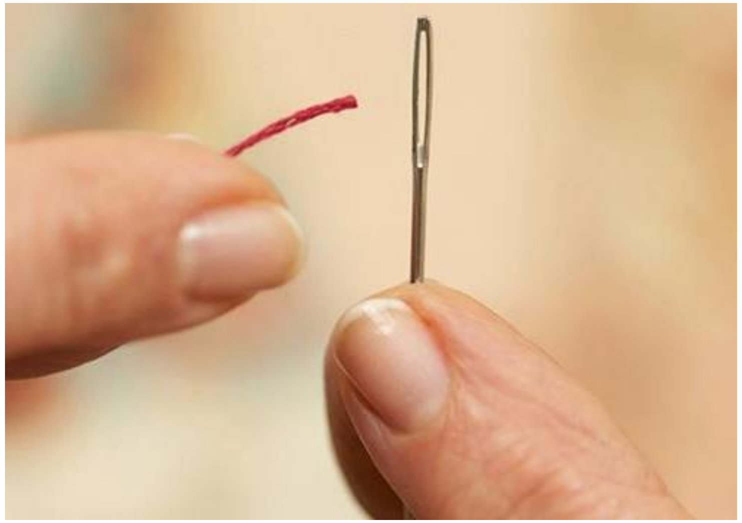
# Challenges

1.




Many tasks and their variations

3.



Precision can be crucial

2.

Current observation is insufficient

4.

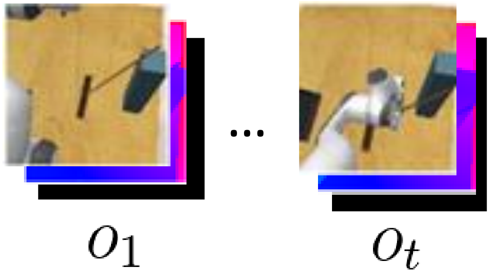
Explicit state recovery is too difficult

# How to address these challenges?

1. Define tasks by language, e.g. Use the broom to brush the dirt into the dustpan

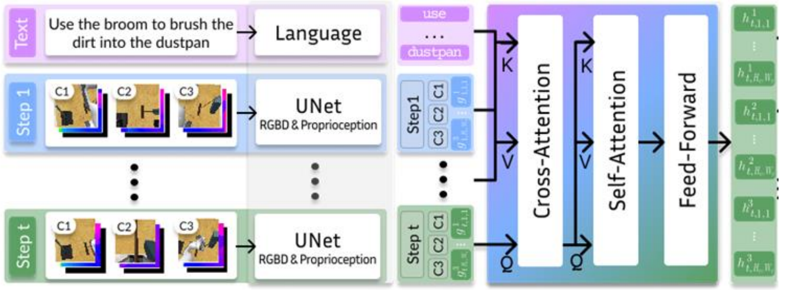
Many tasks and their variations

2. Encode explicit observation history




obse
insufficient

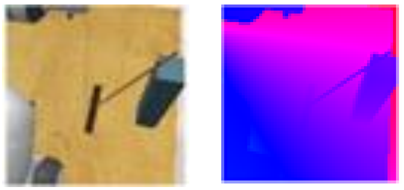
## Transformer




3. Use multi-camera input


Precision can be crucial

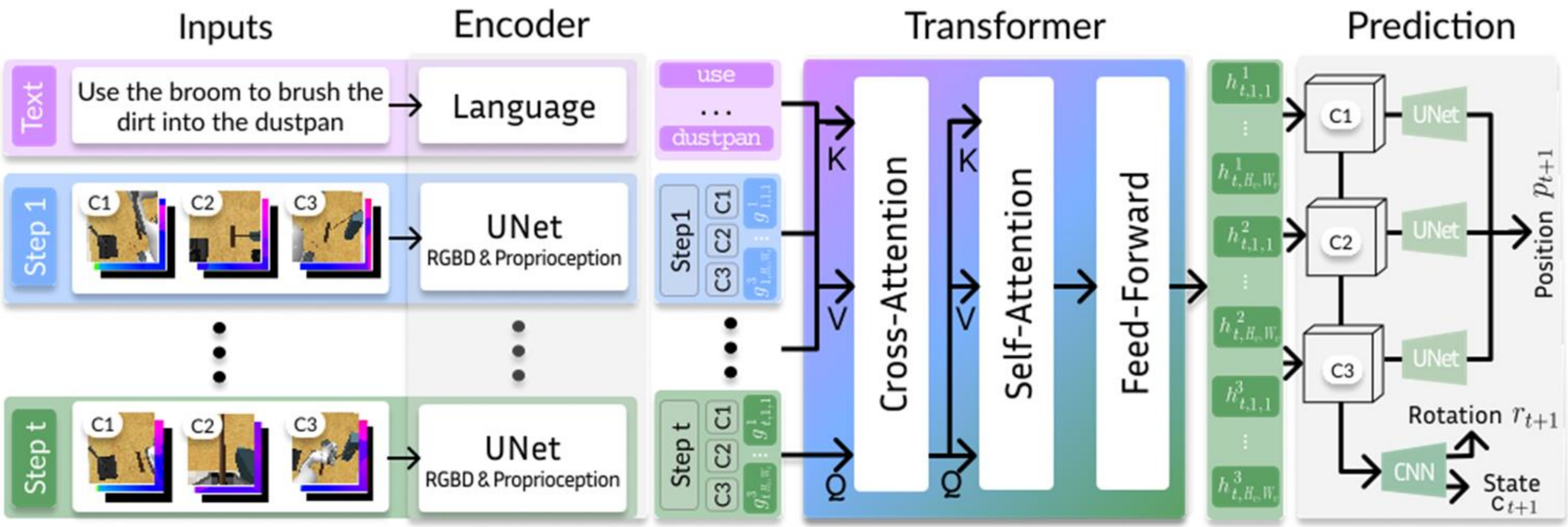
4. Use raw RGB+D for visuomotor policies



state recovery is too difficult

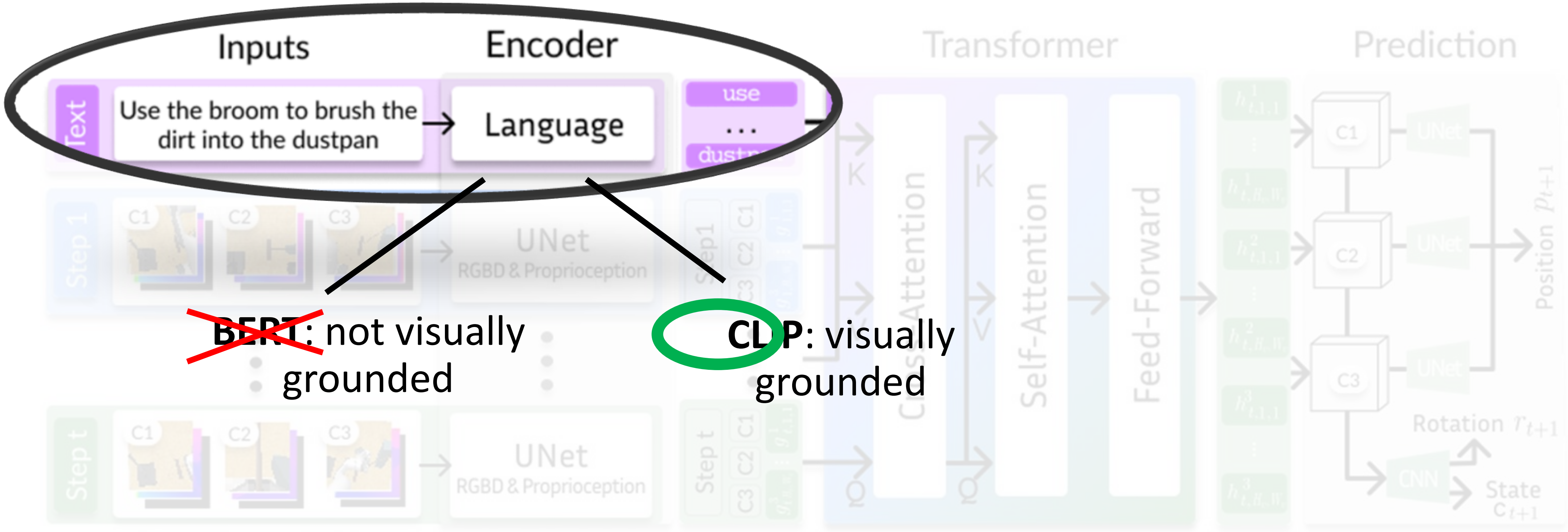
HiveFormer

**H**istory-aware **i**nstruction-conditioned multi-**vie**w trans**former**



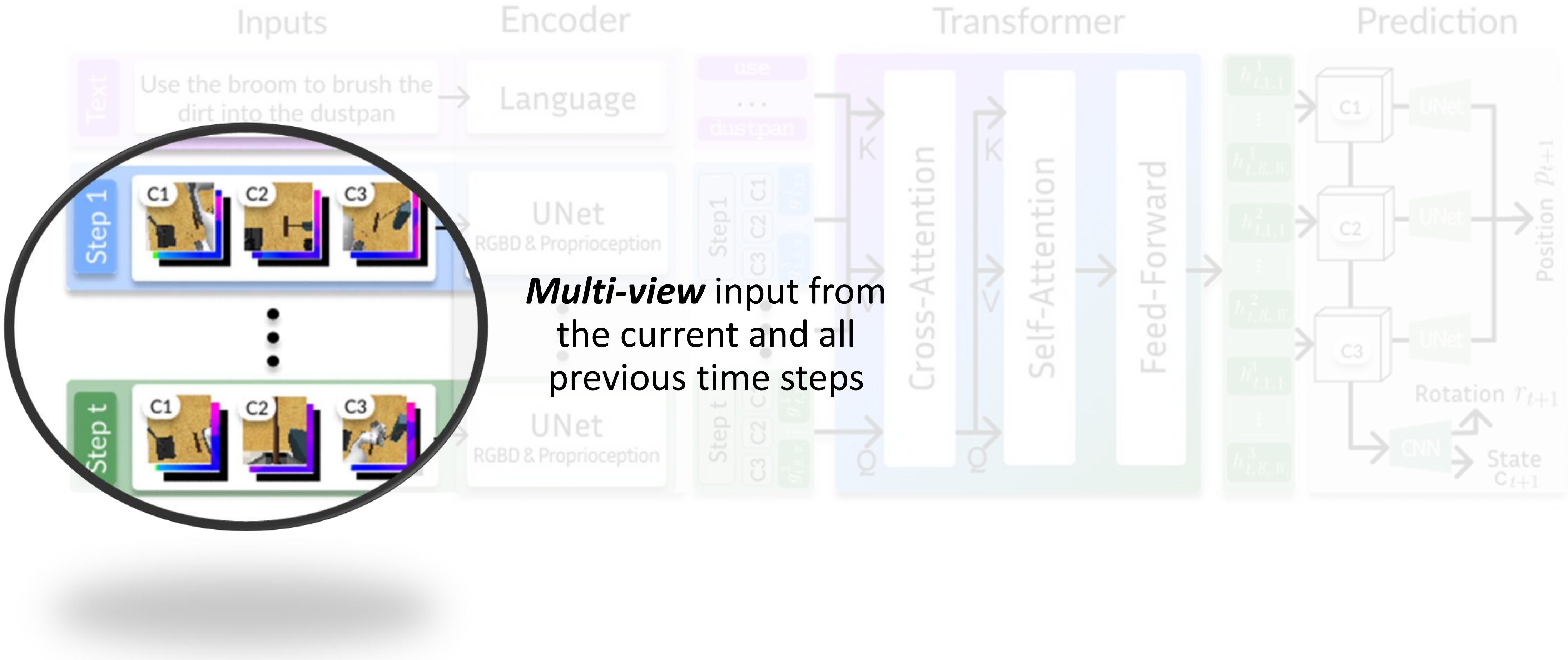
# HiveFormer

## History-aware instruction-conditioned multi-view transformer



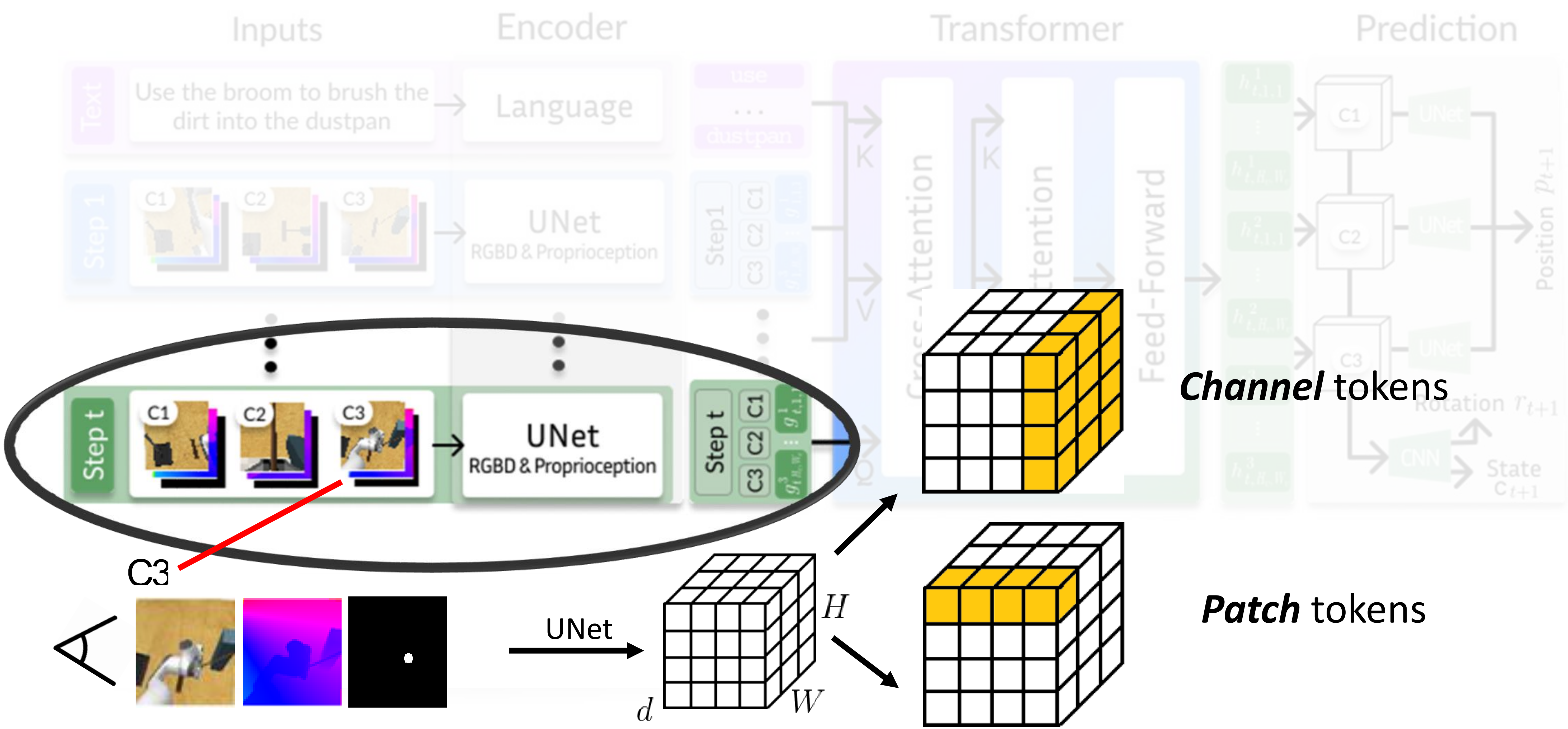
HiveFormer

**H**istory-aware **i**nstruction-conditioned multi-**v**i**e**w trans**former**



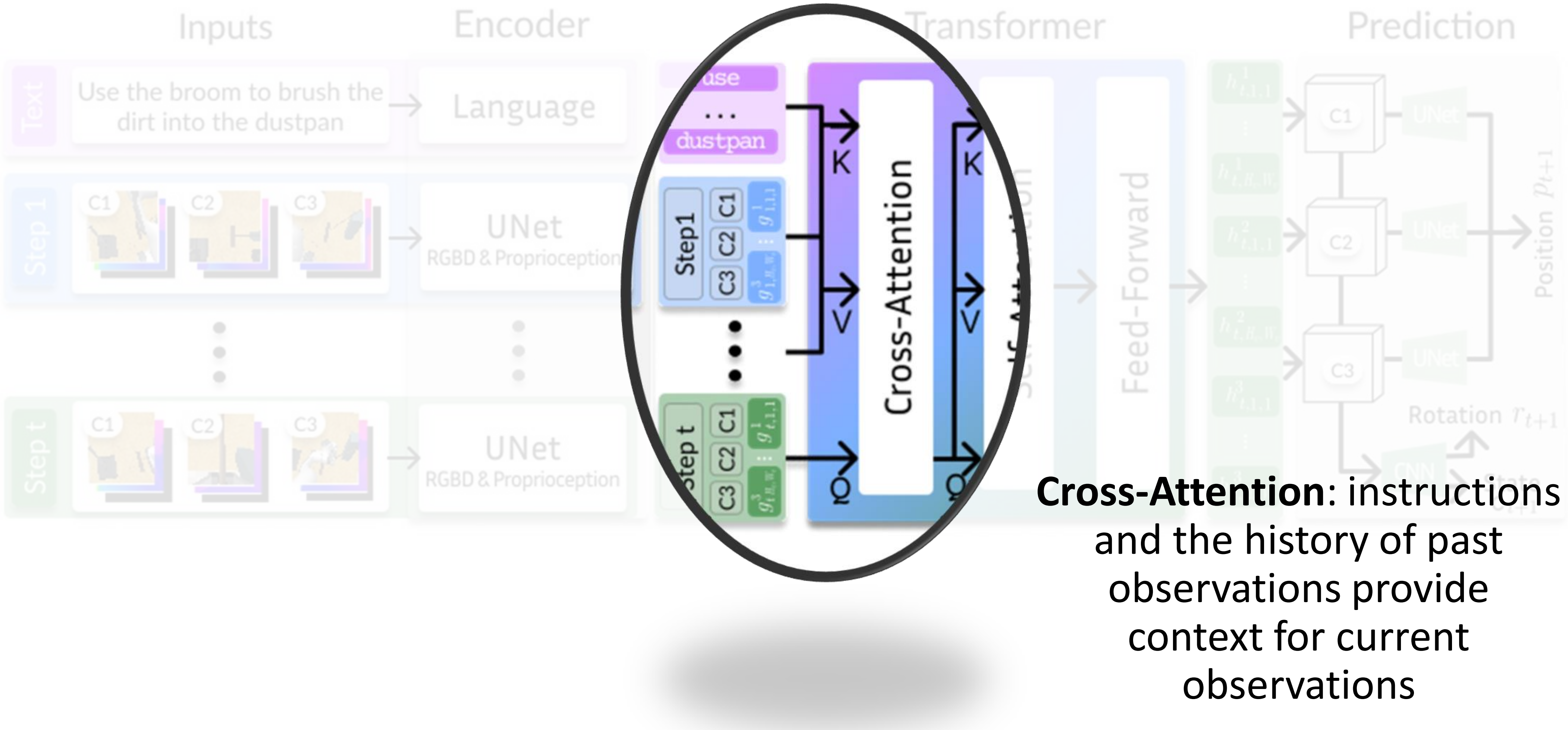
HiveFormer

History-aware instruction-conditioned multi-view transformer



# HiveFormer

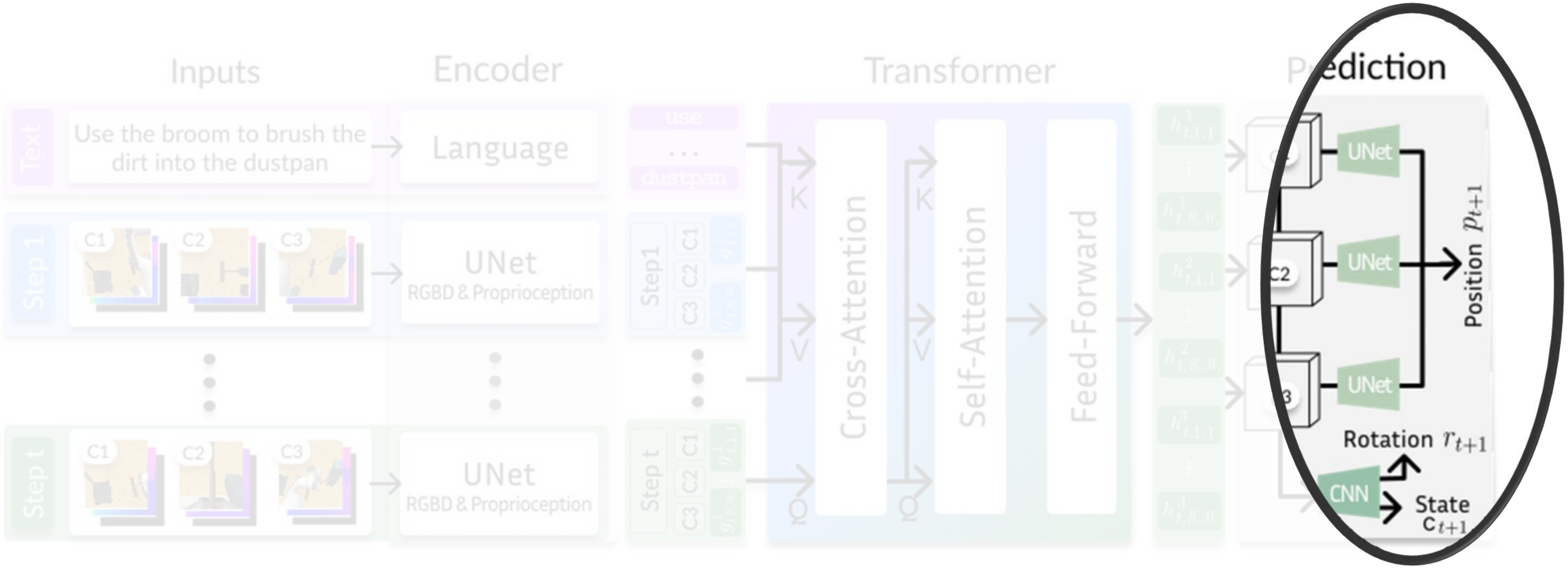
## History-aware instruction-conditioned multi-view transformer




**Cross-Attention**: instructions and the history of past observations provide context for current observations

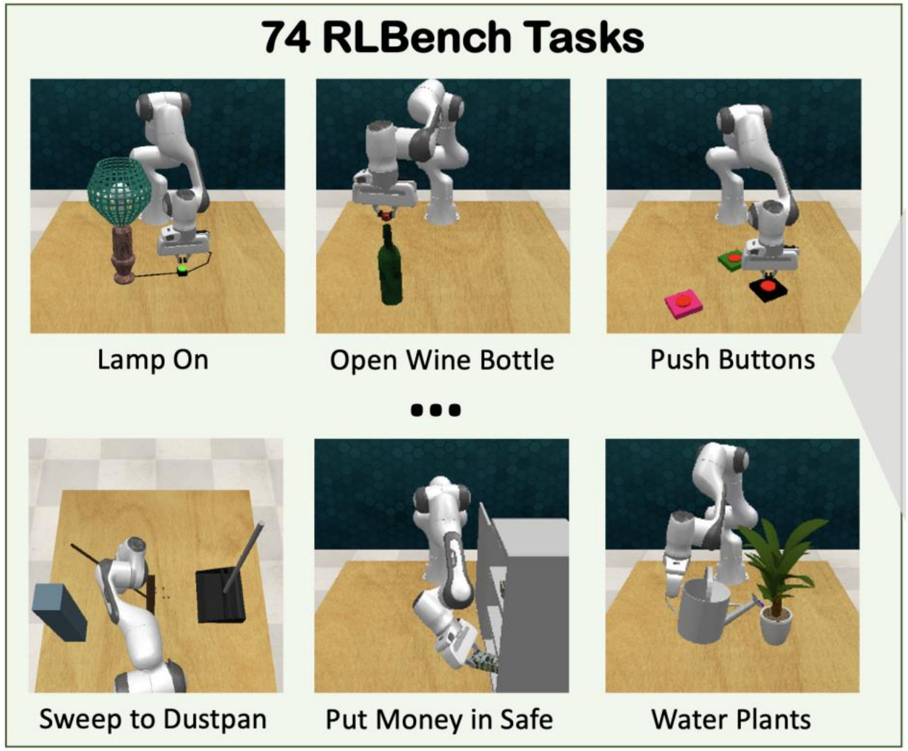
HiveFormer

**H**istory-aware **i**nstruction-conditioned multi-**vie**w trans**former**




**Behavior Cloning** loss for training; Single and Multi-task training

# Evaluation: RLBench tasks




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

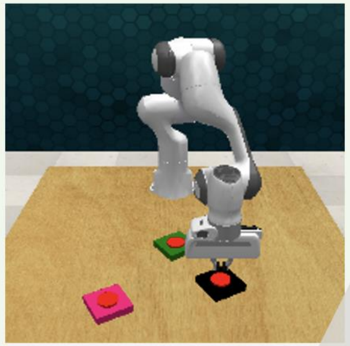
Select 74 tasks we could simulate

Evaluate in single and multi-task settings

(Task text descriptions are not needed)

James, S., Ma, Z., Arrojo, D. R., & Davison, A. J. (2020). RLBench: The robot learning benchmark & learning environment. *IEEE Robotics and Automation Letters*, *5*(2), 3019-3026.

# Evaluation: RLBench task variations

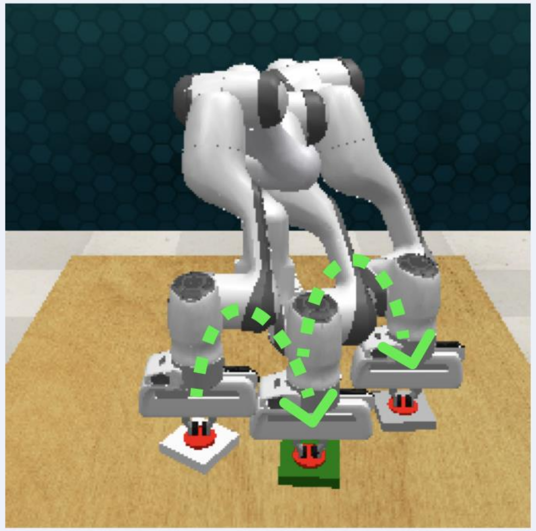


Push Buttons

Water Plants

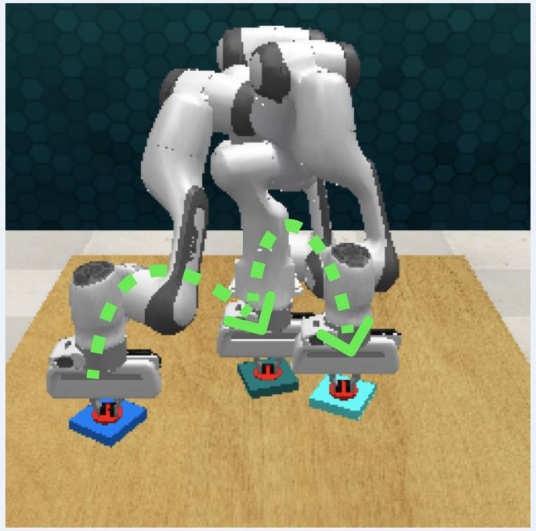
## Variations of a Task

**Training Episode (Seen Variation)**



Press the white button, then push the green button, then push the gray one.

**Testing Episode (Unseen Variation)**



Press the darker blue button, before tapping on the green button and then the lighter blue button.

Unseen sequence of colors during training

Evaluate on *unseen task variations*

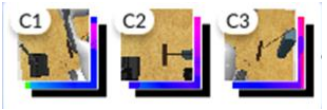
Task text descriptions become crucial

# Results: 10 tasks • Single-task setting

| | Visual Tokens | Point Clouds | Gripper Position | Multi-View | History | Attn | Mask Obs | SR |
|---|---|---|---|---|---|---|---|---|
| R1 | × | × | × | × | × | × | × | $72.9 \pm 4.1$ |
| R2 | Channel | × | × | ✓ | × | Self | × | $73.1 \pm 4.5$ |
| R3 | Channel | ✓ | × | ✓ | × | Self | × | $77.1 \pm 5.8$ |
| R4 | Channel | ✓ | ✓ | ✓ | × | Self | × | $78.1 \pm 5.8$ |
| R5 | Channel | ✓ | ✓ | ✓ | ✓ | Self | × | $81.8 \pm 5.2$ |
| R6 | Channel | ✓ | ✓ | ✓ | ✓ | Self | ✓ | $82.3 \pm 5.3$ |
| R7 | Patch | ✓ | ✓ | ✓ | ✓ | Self | ✓ | $84.4 \pm 6.4$ |
| R8 | Patch | ✓ | ✓ | ✓ | ✓ | Cross | ✓ | $88.4 \pm 4.9$ |

Transformer with multi-view, depth and gripper: +5.2%
w/ vs. w/o history: +3.7%
Patch vs. channel tokens: +2.1%
Cross- vs. Self-Attention: +4%
Overall: +15.5%

# Results: 10 tasks • Single-task setting

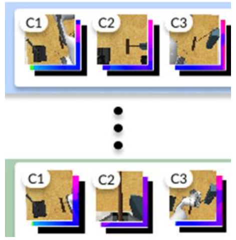



| | Visual Tokens | Point Clouds | Gripper Position | Multi-View | History | Attn | Mask Obs | SR |
|---|---|---|---|---|---|---|---|---|
| R1 | × | × | × | × | × | × | × | $72.9 \pm 4.1$ |
| R2 | Channel | × | × | ✓ | × | Self | × | $73.1 \pm 4.5$ |
| R3 | Channel | ✓ | × | ✓ | × | Self | × | $77.1 \pm 5.8$ |
| R4 | Channel | ✓ | ✓ | ✓ | × | Self | × | $78.1 \pm 5.8$ |
| R5 | Channel | ✓ | ✓ | ✓ | ✓ | Self | × | $81.8 \pm 5.2$ |
| R6 | Channel | ✓ | ✓ | ✓ | ✓ | Self | ✓ | $82.3 \pm 5.3$ |
| R7 | Patch | ✓ | ✓ | ✓ | ✓ | Self | ✓ | $84.4 \pm 6.4$ |
| R8 | Patch | ✓ | ✓ | ✓ | ✓ | Cross | ✓ | $88.4 \pm 4.9$ |

+5.2 %

Transformer with multi-view, depth and gripper: +5.2%
w/ vs. w/o history: +3.7%
Patch vs. channel tokens: +2.1%
Cross- vs. Self-Attention: +4%
Overall: +15.5%

# Results: 10 tasks • Single-task setting

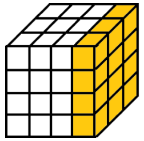


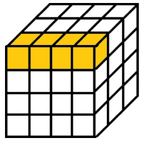

| | Visual Tokens | Point Clouds | Gripper Position | Multi-View | History | Attn | Mask Obs | SR |
|---|---|---|---|---|---|---|---|---|
| R1 | × | × | × | × | × | × | × | 72.9 ± 4.1 |
| R2 | Channel | × | × | ✓ | × | Self | × | 73.1 ± 4.5 |
| R3 | Channel | ✓ | × | ✓ | × | Self | × | 77.1 ± 5.8 |
| R4 | Channel | ✓ | ✓ | ✓ | × | Self | × | 78.1 ± 5.8 |
| R5 | Channel | ✓ | ✓ | ✓ | ✓ | Self | × | 81.8 ± 5.2 |
| R6 | Channel | ✓ | ✓ | ✓ | ✓ | Self | ✓ | 82.3 ± 5.3 |
| R7 | Patch | ✓ | ✓ | ✓ | ✓ | Self | ✓ | 84.4 ± 6.4 |
| R8 | Patch | ✓ | ✓ | ✓ | ✓ | Cross | ✓ | 88.4 ± 4.9 |

+3.7 %

Transformer with multi-view, depth and gripper: +5.2%
w/ vs. w/o history: +3.7%
Patch vs. channel tokens: +2.1%
Cross- vs. Self-Attention: +4%
Overall: +15.5%

# Results: 10 tasks • Single-task setting





| | Visual Tokens | Point Clouds | Gripper Position | Multi-View | History | Attn | Mask Obs | SR |
|---|---|---|---|---|---|---|---|---|
| R1 | × | × | × | × | × | × | × | 72.9 ± 4.1 |
| R2 | Channel | × | × | ✓ | × | Self | × | 73.1 ± 4.5 |
| R3 | Channel | ✓ | × | ✓ | × | Self | × | 77.1 ± 5.8 |
| R4 | Channel | ✓ | ✓ | ✓ | × | Self | × | 78.1 ± 5.8 |
| R5 | Channel | ✓ | ✓ | ✓ | ✓ | Self | × | 81.8 ± 5.2 |
| R6 | Channel | ✓ | ✓ | ✓ | ✓ | Self | ✓ | 82.3 ± 5.3 |
| R7 | Patch | ✓ | ✓ | ✓ | ✓ | Self | ✓ | 84.4 ± 6.4 |
| R8 | Patch | ✓ | ✓ | ✓ | ✓ | Cross | ✓ | 88.4 ± 4.9 |

+2.1 %

Transformer with multi-view, depth and gripper: +5.2%

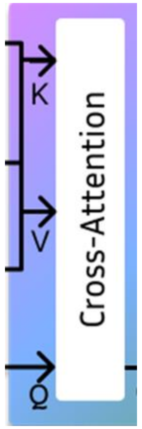
w/ vs. w/o history: +3.7%

Patch vs. channel tokens: +2.1%

Cross- vs. Self-Attention: +4%

Overall: +15.5%

# Results: 10 tasks • Single-task setting

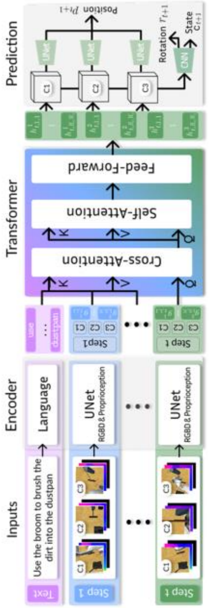



| | Visual Tokens | Point Clouds | Gripper Position | Multi-View | History | Attn | Mask Obs | SR |
|---|---|---|---|---|---|---|---|---|
| R1 | × | × | × | × | × | × | × | 72.9 ± 4.1 |
| R2 | Channel | × | × | ✓ | × | Self | × | 73.1 ± 4.5 |
| R3 | Channel | ✓ | × | ✓ | × | Self | × | 77.1 ± 5.8 |
| R4 | Channel | ✓ | ✓ | ✓ | × | Self | × | 78.1 ± 5.8 |
| R5 | Channel | ✓ | ✓ | ✓ | ✓ | Self | × | 81.8 ± 5.2 |
| R6 | Channel | ✓ | ✓ | ✓ | ✓ | Self | ✓ | 82.3 ± 5.3 |
| R7 | Patch | ✓ | ✓ | ✓ | ✓ | Self | ✓ | 84.4 ± 6.4 |
| R8 | Patch | ✓ | ✓ | ✓ | ✓ | Cross | ✓ | 88.4 ± 4.9 |

+4
%

Transformer with multi-view, depth and gripper: +5.2%
w/ vs. w/o history: +3.7%
Patch vs. channel tokens: +2.1%
Cross- vs. Self-Attention: +4%
Overall: +15.5%

# Results: 10 tasks • Single-task setting

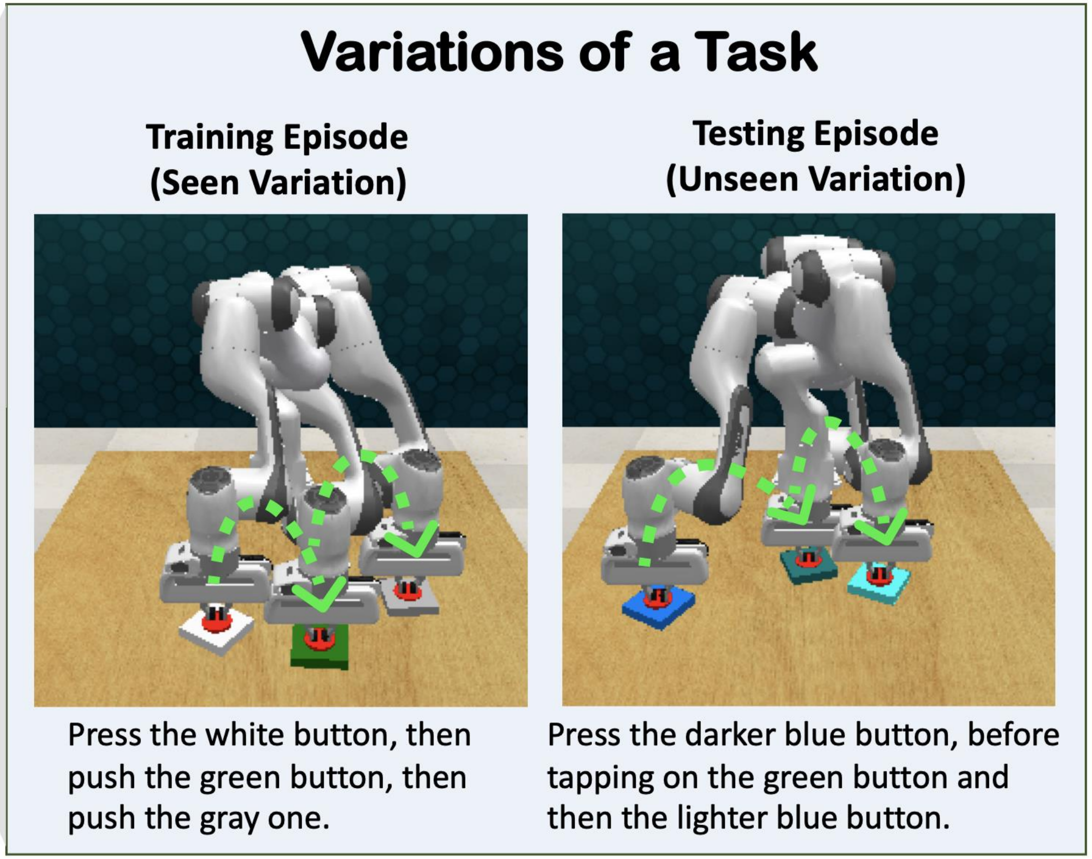



| | Visual Tokens | Point Clouds | Gripper Position | Multi-View | History | Attn | Mask Obs | SR |
|---|---|---|---|---|---|---|---|---|
| R1 | × | × | × | × | × | × | × | 72.9 ± 4.1 |
| R2 | Channel | × | × | ✓ | × | Self | × | 73.1 ± 4.5 |
| R3 | Channel | ✓ | × | ✓ | × | Self | × | 77.1 ± 5.8 |
| R4 | Channel | ✓ | ✓ | ✓ | × | Self | × | 78.1 ± 5.8 |
| R5 | Channel | ✓ | ✓ | ✓ | ✓ | Self | × | 81.8 ± 5.2 |
| R6 | Channel | ✓ | ✓ | ✓ | ✓ | Self | ✓ | 82.3 ± 5.3 |
| R7 | Patch | ✓ | ✓ | ✓ | ✓ | Self | ✓ | 84.4 ± 6.4 |
| R8 | Patch | ✓ | ✓ | ✓ | ✓ | Cross | ✓ | 88.4 ± 4.9 |

+15.5 %

Transformer with multi-view, depth and gripper: +5.2%
w/ vs. w/o history: +3.7%
Patch vs. channel tokens: +2.1%
Cross- vs. Self-Attention: +4%
Overall: +15.5%

# 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 |

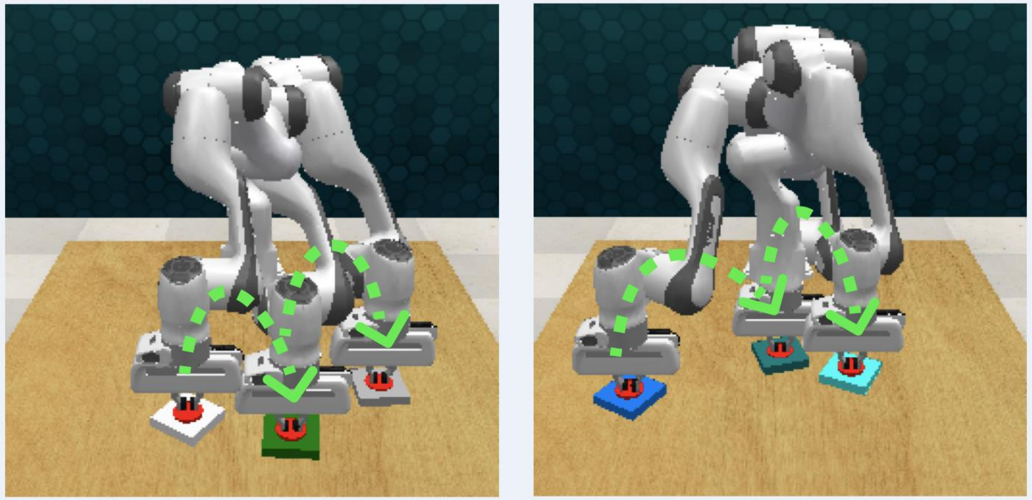
Generalization to unseen variations

Generalization to natural language extractions

# Results: Task variations

## Variations of a Task

| **Training Episode (Seen Variation)** | **Testing Episode (Unseen Variation)** |
|---|---|



| Press the white button, then push the green button, then push the gray one. | Press the darker blue button, before tapping on the green button and then the lighter blue button. |
|---|---|

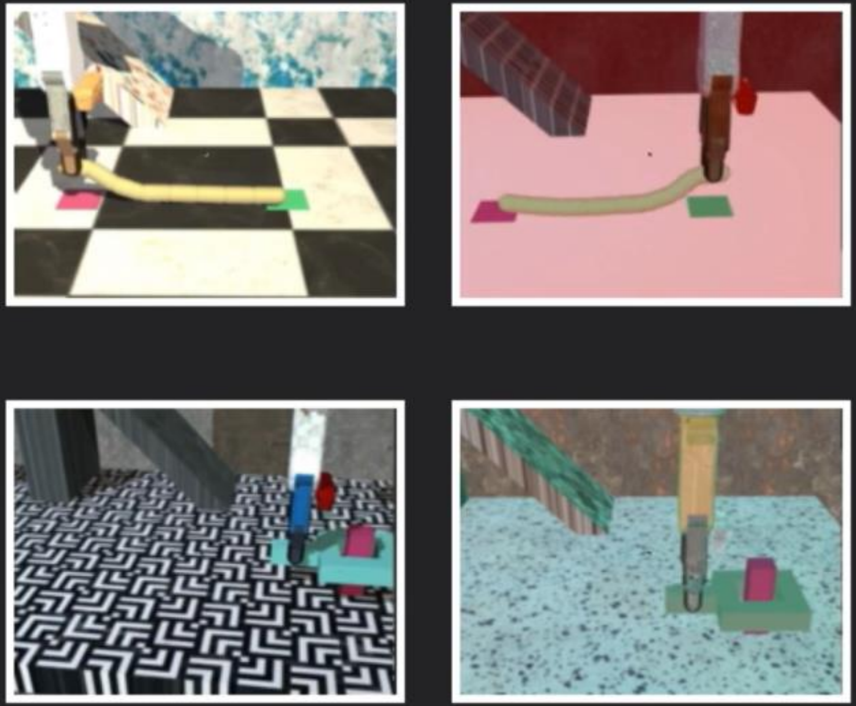
| # Demos Per Variation | Instr. | Push Buttons | | | Tower | | |
|---|---|---|---|---|---|---|---|
| | | Seen | Unseen | | Seen | Unseen | |
| | | Synt. | Synt. | Real | Synt. | Synt. | Real |
| 10 | Seq. | 96.4 | 71.1 | 65.7 | 71.6 | 49.8 | 19.4 |
| 50 | Seq. | 99.4 | 83.1 | 70.9 | 74.3 | 52.1 | 20.6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

Generalization to unseen variations

Generalization to natural language expressions

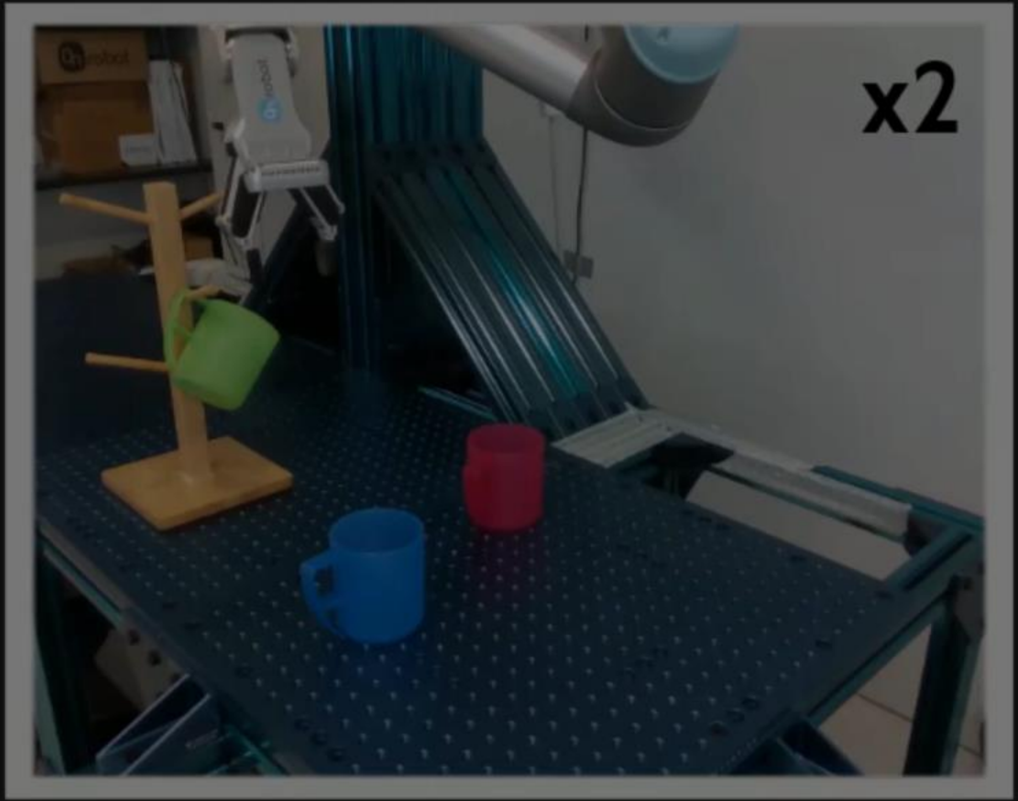
# Domain randomization

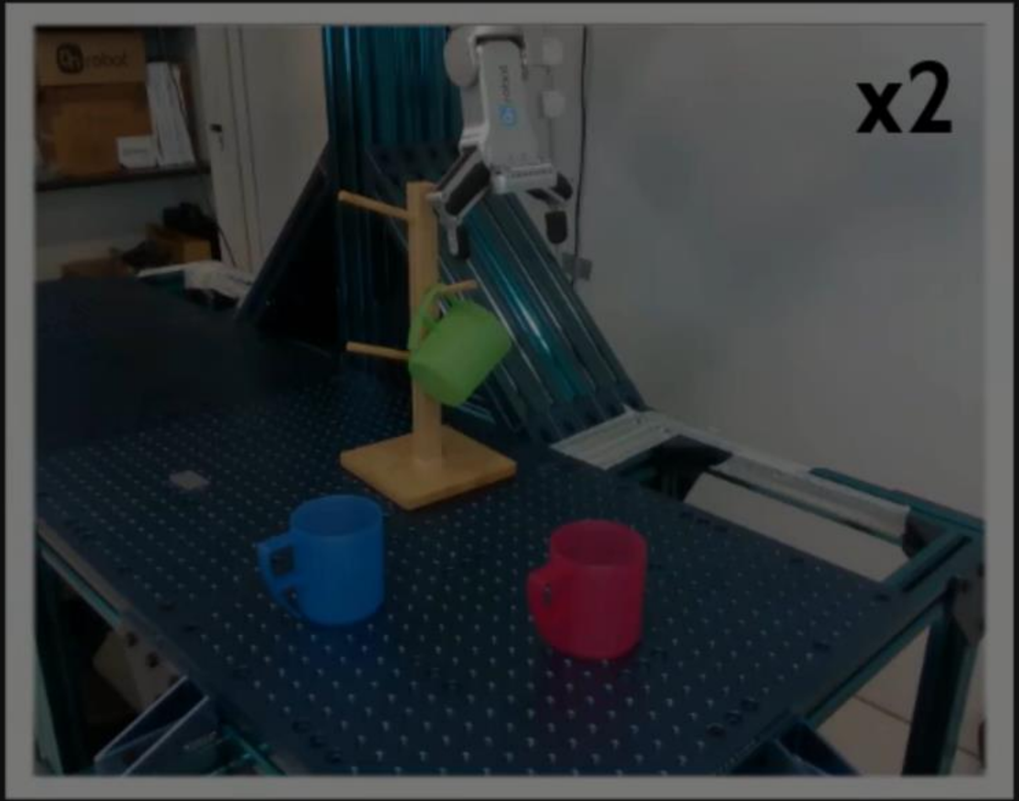
Training: simulated scenes



Testing: real scenes

# Experiments for Hang Mug Task



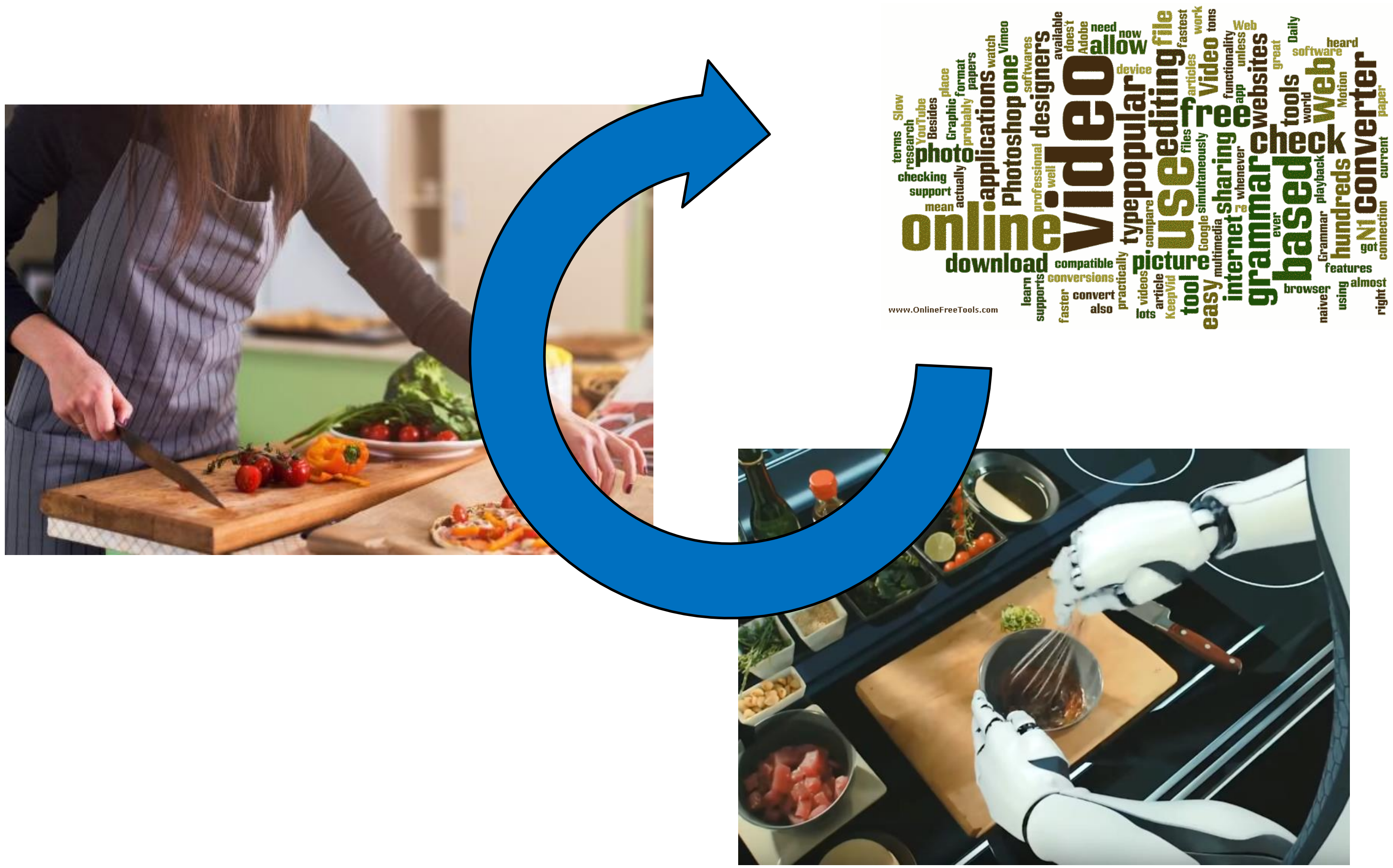
# Vision should be grounded in real actions

- Vision requires models of physics and actions in the real world

- Robotics requires models of vision and perception

# Vision, language and robotics

Goal: learn Large Embodied Vision-Language Models (LEVLM)



# Thanks to my collaborators and students


Cordelia Schmid


Josef Sivic


Jean Ponce


Francis Bach


J.-P. Laumonde


J. Carpentier


Andrew Zisserman


Aloysha Efros


Michael Black


Shizhe Chen


M. Tapaswi


Vijay Kumar


Karteek Alahari


Gul Varol


Yana Hasson


Antoine Yang

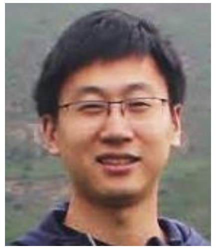

E. Chane-Sane


Antoine Miech


P.-L. Guhur


Robin Strudel


R. Garcia Pinel


Zerui Chen


J.-B. Alayrac


I. Kalevatkh


A. Pashevich


Q. Le Lidec


Alaa El-Nouby


M. Futeral-Peter


D. Zhukov


Vincent Delaitre


G. Seguin


Guilhem Cheron


Piotr Bojanowski

MOHAMED BIN ZAYED
UNIVERSITY OF
ARTIFICIAL INTELLIGENCE



# Ranked in the Top 20 globally in AI, CV, ML and NLP

READ MORE | RESEARCH | SUSTAINABILITY

MOHAMED BIN ZAYED
UNIVERSITY OF
ARTIFICIAL INTELLIGENCE

Login Careers Quick links EN AR

About Study Research Innovate News & events

# Building a new lab for Embodied and Language-Aware Visual Models

Ranked in the Top 20 globally in AI, CV, ML and NLP

READ MORE | RESEARCH | SUSTAINABILITY