SMU
SINGAPORE MANAGEMENT UNIVERSITY

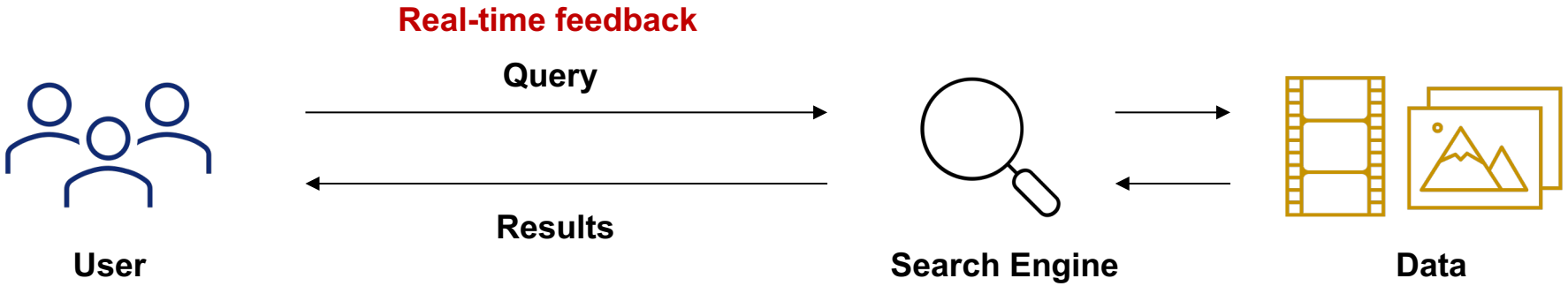
School of
Information Systems

Reinforcement Learning Enhanced Interactive Video Retrieval

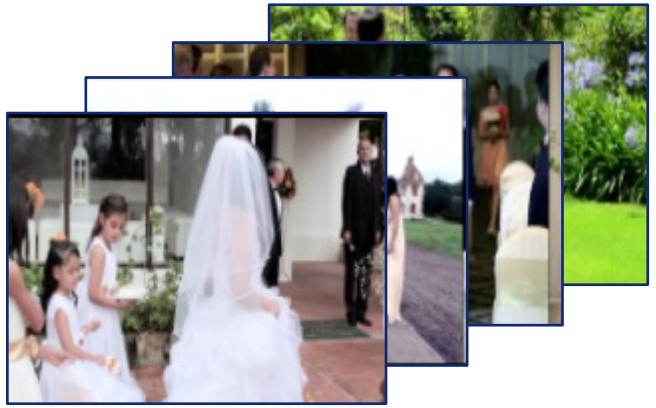
MA Zhixin
Ph.D. Student
School of Computing and Information System
Singapore Management University

What is Interactive Video Retrieval?

- Automatic *vs* Interactive video retrieval?



A wedding party which is grouped around the bride and groom, bride and groom are walking and guests are following them.



Visual Known-Item Search

- **Visual known-item search (V-KIS)**: Assume that you watched this video before and want to find back it from video collection.

A V-KIS Query:

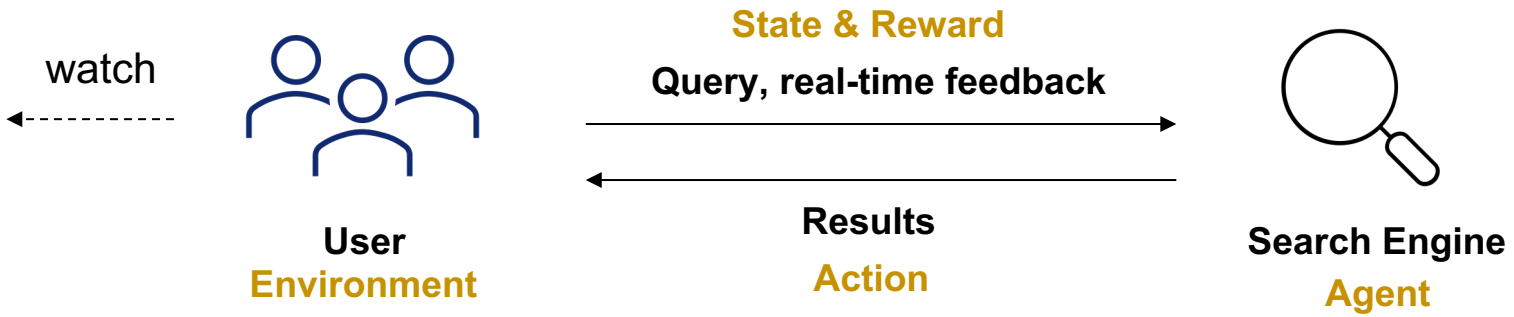


Table. Dataset statistic showing the number of video shots, duration and release year

Dataset	Clip	Duration	Year
$V3C_1$	1,082,659	1,000 h	2019
$V3C_2$	1,425,454	1,300 h	2022
Total	2,508,113	2,300 h	-

Video Browser Showdown

- **Video browser showdown (VBS)**: A **live** video search competition, where teams evaluate and demonstrate the efficiency of their video retrieval tools on a **shared dataset**

Visual Know-Item Search

- Trial-and-error is inefficient
- Inspection is tedious and time-consuming
- Sub-optimal results harm the search
- Concept selection
- …

Interactive Video Corpus Moment Retrieval using Reinforcement Learning

An exploration to improve interactive video retrieval using reinforcement learning

“Finding needle from a haystack”

A V-KIS query from Video Browsing Showdown (VBS) 2022

Figure. An example of search target.

Query:

People climb on a mountain

↓

People climb on a mountain next to bridge and river

Snow mountain ✗

Forest ✗

Similar

“Finding needle from a haystack”

Issues of the current interaction:

- trial-and-error is inefficient
- inception is tedious and time consuming
- sessions are independent

Motivation: an interactive video search framework

- allow user provide feedback to update query
- navigate user from a sub-optimal result to the target

Figure. A representative scene in the search target.

Figure. Search results after adding “bridge” and “river” to the query.

Related Work

- Single video moment retrieval
- Domain-specific dialogue-based interactive image retrieval

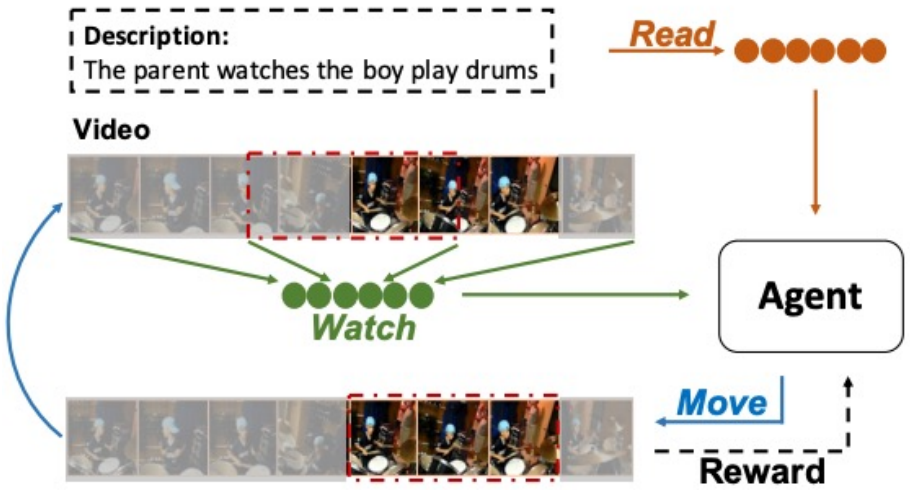


Figure. A framework of single video moment retrieval. [1]

Figure. An example of dialogue-based interactive image retrieval. [2]

[1] Read, watch, and move: Reinforcement learning for temporally grounding natural language descriptions in videos, He et al., AAAI 2019
[2] Dialog-based Interactive Image Retrieval. Guo et al., NIPS 2018

Problem Statement

- Users provide feedback $\mathrm{f_t}$ to candidate
- Navigate from starting moment m_0 to search target m^*

  - Graph of moment: $\mathcal{G}$
    - Pairwise similarity including semantic & visual feature
  - Graph walking as Markov Decision Process (MDP)
    - State $\mathcal{S}$, Action Space $\mathcal{A}$, Reward Function $\mathcal{R}$, Transition Probability $\mathcal{P}$

Rachel explains to her dad on the phone why she can't marry her fiancé.

Initial Query: q_0

start moment: m_0

user feedback: $\mathrm{f_t}$

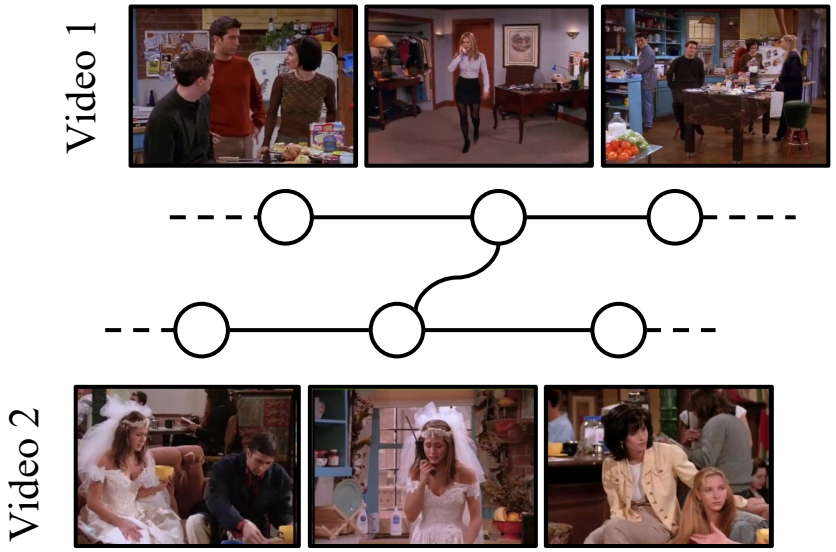
- not in office
- in wedding dress

search target (m^*)

Problem Statement

Link video shots as a graph

- Successive shots
- Shots with high similarity
  - Threshold, …



(a)

- Connectivity
- Density
- …

- Concept detection
- Visual feature
- Script TF-IDF

Rachel, phone, call, …
office, suit, …

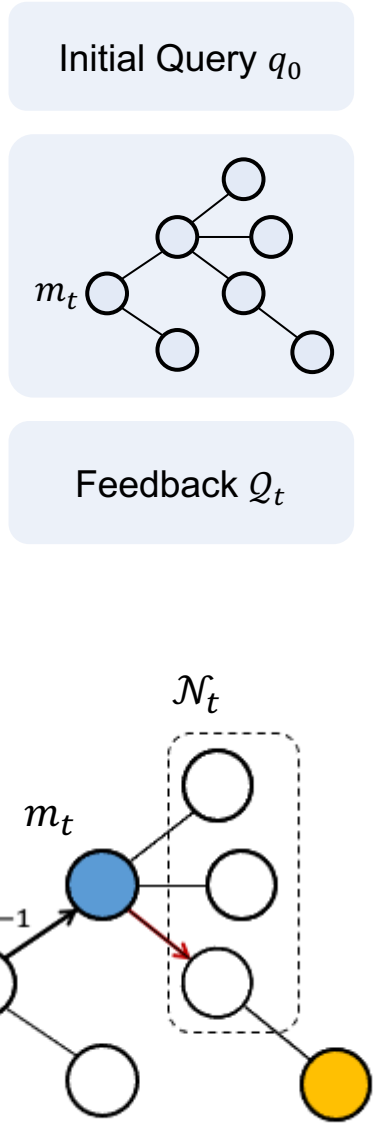
Rachel, phone, call, …
wedding, dress, …

(b)

Graph Walking as MDP

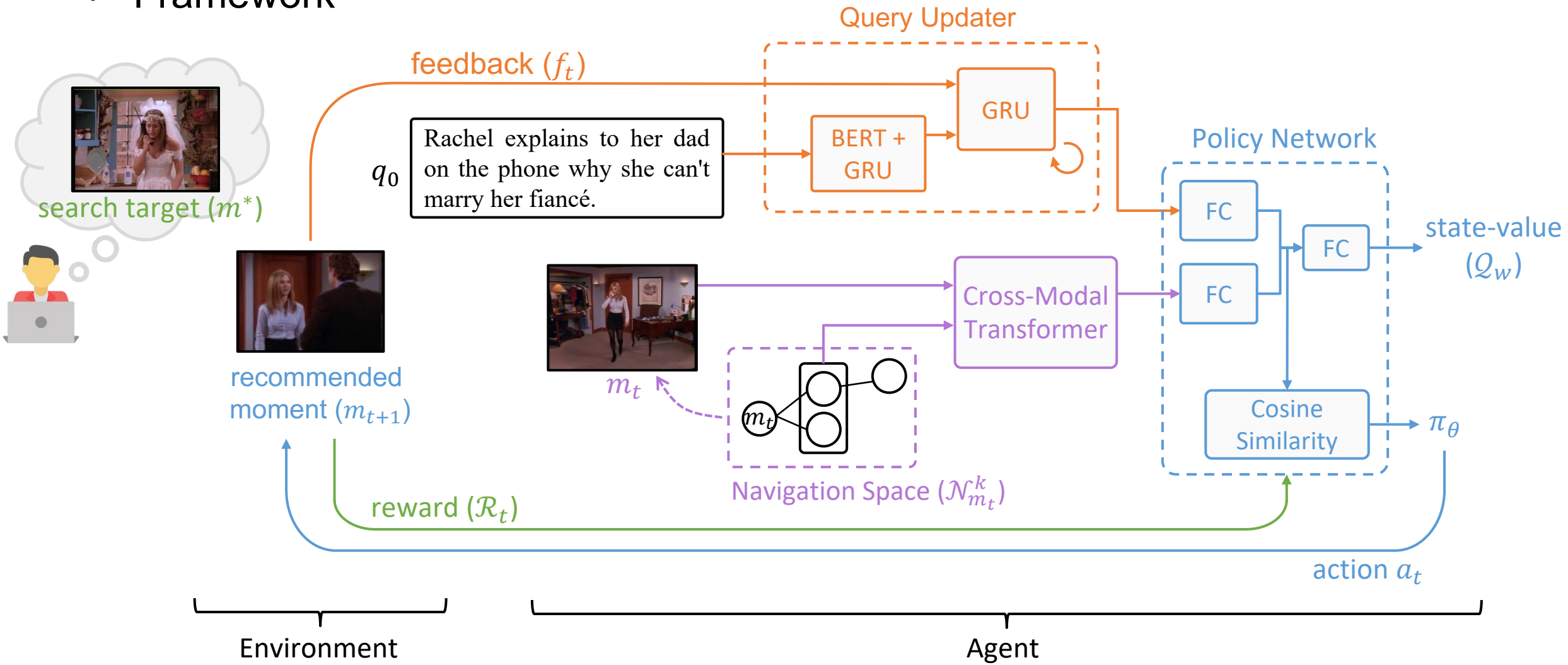
Markov Decision Process

- State Space $\mathcal{S}$
  - $s_t = (q_0, m_t, f_t)$
- Action Space $\mathcal{A}$
  - Neighbors within k steps
- Reward Function $\mathcal{R}$
  - $r_t = \begin{cases} \frac{1}{2^{d_{t+1}}} - \phi \cdot t, & d_t > d_{t+1} \\ -\phi \cdot t, & d_t = d_{t+1} \\ -\frac{1}{2} - \phi \cdot t, & d_t < d_{t+1} \end{cases}$
- Transition Probability $\mathcal{P}$



Methodology

- Framework



User Simulator

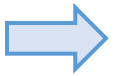
- Keyword-level feedback
  1. Detect semantic concepts for the target and current shots
  2. Calculate pair-wise probability difference
  3. Sample keyword from the distribution to add or remove keywords

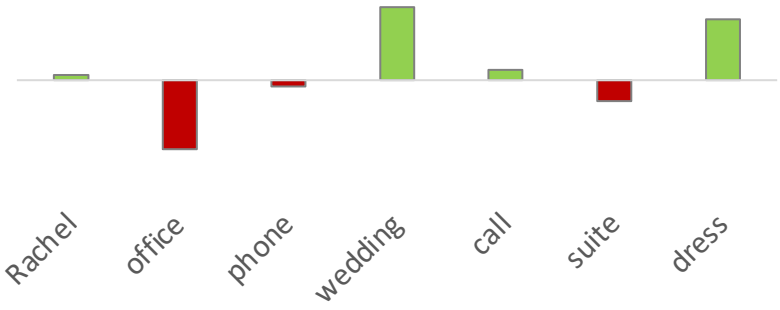
target moment m^*

Rachel, phone, call, …
wedding, dress, …

current moment m_t

Rachel, phone, call, …
office, suit, …





Dataset

- TV Retrieval [1] and DiDeMo [2]

Table 1. Dataset statistic showing the number of videos and queries in different splits.

Dataset	#Video			#Query		
	Train	Val	Test	Train	Val	Test
TVR	17,435	2,179	-	87,175	10,895	-
DiDeMo	8,395	1,065	1,004	32,624	4,160	3,982

Table 2. Dataset statistic showing the number of graph nodes, edges and average degree.

Dataset	#Node	#Edge	Avg Degree
TVR	147,985	294,395	3.98
DiDeMo	114,923	263,609	4.59

[1] TVR: A Large-Scale Dataset for Video-Subtitle Moment Retrieval, Lei et al., ECCV 2020
[2] Localizing Moments in Video with Temporal Language, Hendricks et al., EMNLP 2018

Model Training

Simulation | Real Scenario | Human Performance

1. Initialize $\pi_{\theta_\pi}(a|s)$ and $Q_{\theta_Q}(a,s)$
2. Sample a set of m_0 ● for each query q and m^* ●
3. For each query q:
4. For $t = 0 \ldots T$:
5. estimate $\pi_\theta(a_i|s_t)$ and state value $Q(a_i, s_t)$
6. take a_t, receive r_t and s_{t+1}
7. $s_t \leftarrow s_{t+1}$
8. update parameter θ_π and θ_Q

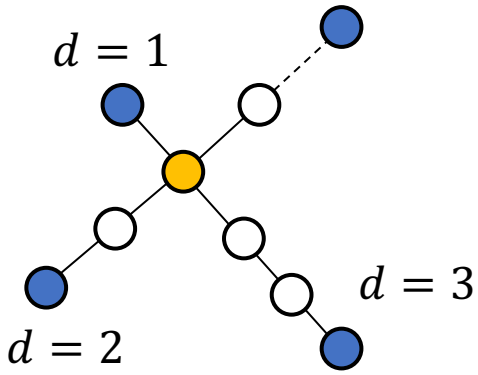


Figure. Illustration of node sampling.

$$L_{triplet} = max(0, c + u^+ - u^-)$$

$$u^{\{+/-\}} =< FC_m(\mathbf{m}^{\{+/-\}}) \cdot FC_q(\mathbf{q}) >$$

Imitation Learning

$$L_{policy} = -\sum_t log\, \pi_\theta(a_t|s_t)(R_t - Q_w(s_t))$$

$$L_{mse} = \sum_t (Q_w(s_t) - R_t)^2$$

Advantage Actor Critic (A2C)

Experiments

Simulation | Real Scenario | Human Performance

- Model Comparison
  - Random: randomly select a moment at each time
  - No-Feedback: greedily select actions w/o query updating
  - Imitation: optimize policy using supervised learning
  - Ours: optimize policy using both supervised learning and reinforcement learning (i.e., A2C)

Table. Model comparison of recall@1 on TVR and DiDeMo dataset

Model	TVR				DiDeMo			
	d_0=1	d_0=2	d_0=3	d_0=4	d_0=1	d_0=2	d_0=3	d_0=4
Random	0.058 ±0.001	0.044 ±0.001	0.024 ±0.001	0.007 ±0.001	0.093 ±0.004	0.078 ±0.003	0.053 ±0.005	0.023 ±0.001
No-Feedback	0.489 ±0.011	0.475 ±0.013	0.442 ±0.012	0.204 ±0.008	0.259 ±0.004	0.242 ±0.001	0.229 ±0.005	0.121 ±0.002
Imitation	0.514 ±0.011	0.485 ±0.010	**0.448** ±0.011	0.210 ±0.006	0.275 ±0.005	0.261 ±0.009	0.242 ±0.008	0.132 ±0.006
Ours	**0.534** ±0.016	**0.495** ±0.002	0.446 ±0.005	**0.216** ±0.003	**0.296** ±0.003	**0.275** ±0.002	**0.262** ±0.009	**0.141** ±0.003

Does the model work in real-scenario?

Experiment

Simulation | Real Scenario | Human Performance

Does the model work in real-scenario?

- Initialize the start moment m_0 with SOTA automatic VCMR models
  - HERO [1] and CONQUER [2]
  - test whether our model can find the clips which are deeply hidden in the ranked lists

depth of target in a ranked list →

	TVR				DiDeMo			
	(10-50]	(50-100]	(100-200]	>200	(10-50]	(50-100]	(100-200]	>200
HERO	2179 (14%)	822 (9%)	785 (6%)	328 (5%)	683 (9%)	379 (5%)	396 (4%)	886 (3%)
CONQUER	2049 (14%)	681 (9%)	416 (9%)	98 (8%)	823 (8%)	415 (7%)	388 (3%)	591 (3%)

Our model can locate the clips ranked low by the SOTA VCMR models

[1] HERO: Hierarchical Encoder for Video+Language Omni-representation Pre-training, Li et al., EMNLP 2020
[2] CONQUER: Contextual Query-aware Ranking for Video Corpus Moment Retrieval, Hou et al., MM 2021

Experiment

Simulation | Real Scenario | Human Performance

Does the user simulator behave like a human user?

- Replace user simulator with a human user to test the effectiveness of user simulator
- 50 queries which cannot be retrieved by the user simulator

Table. Human performance of identifying 50 search targets that cannot be retrieved by a user simulator.

Interaction Setting	Recall@1	#Keywords	#Iteration
1 keyword 1 shot	18%	1	3.22
5 keywords 1 shot	18%	4.2	3.08
5 keywords 5 shots	44%	4.1	2.5

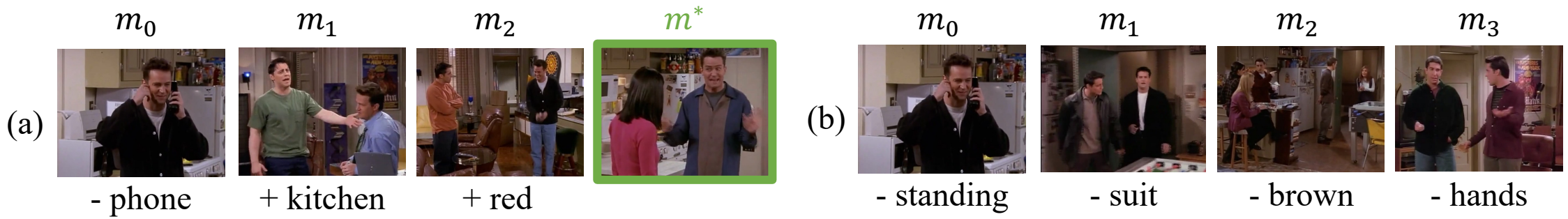


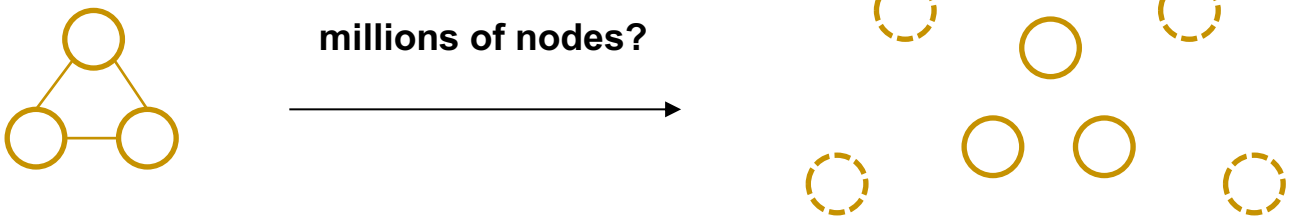
Figure. Example comparing how (a) human and (b) user simulator interact with the agent for the query "Chandler shakes his hands in the air while standing in the kitchen".

Future Work

- Extend the simulator to more vividly mimic human behavior

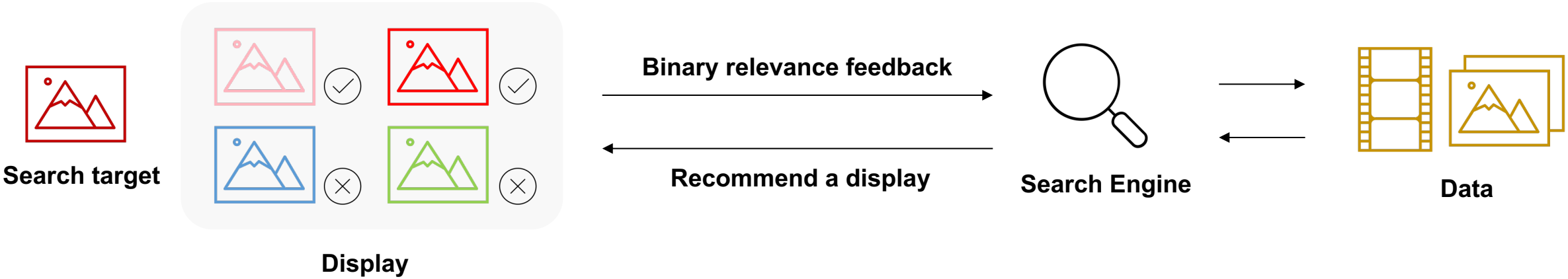
- She should wear a wedding dress
- She is not in the office
- …

- Adapt the user simulator to larger and more general datasets



Ongoing work: Relevance Feedback

- Bayesian-based relevance feedback



- Is the recommendation user-friendly?
- Is the feedback effective?
- …

Thanks