SEOUL NATIONAL UNIV.
VISION & LEARNING

YAHOO! RESEARCH

tGIF QA

Toward Spatio-Temporal Reasoning in Visual Question Answering

Yunseok Jang† Yale Song‡ Youngjae Yu† Youngjin Kim† Gunhee Kim†

Seoul National University† Yahoo! Research‡

Code and Dataset are available at http://vision.snu.ac.kr/projects/tgif-qa

CVPR July 21-26 HONOLULU 2017

Motivation

- Significant progress in **image-based** VQA with various datasets

DAQUAR [Malinowski et al., *NIPS* 2014]

VQA [Antol et al., *ICCV* 2015]

Visual Madlibs [Yu et al., *ICCV* 2015]

COCO-QA [Ren et al., *NIPS* 2015]

FM-IQA [Gao et al., *NIPS* 2015]

Visual7W [Zhu et al., *CVPR* 2016]

How about VQA tasks on videos?

- A few datasets use movie as data source for video VQA

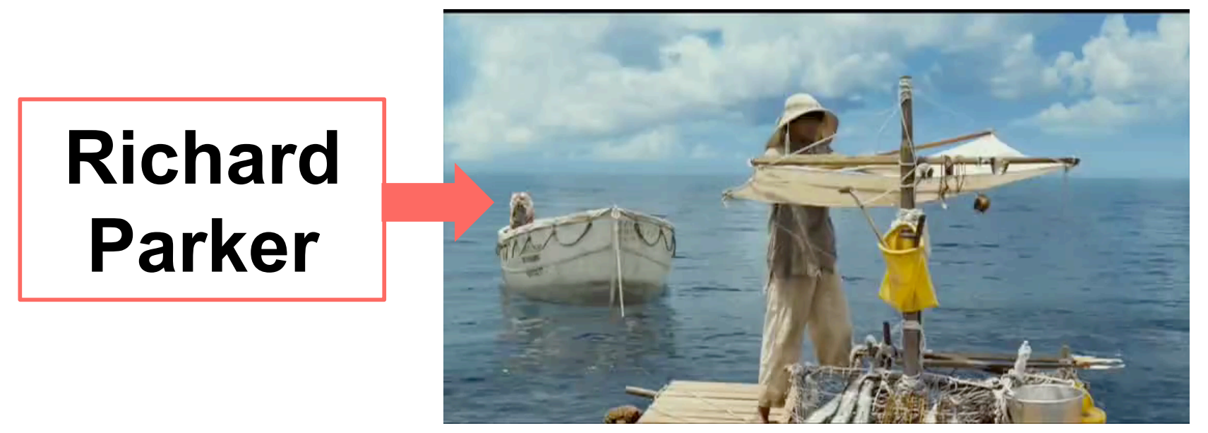
MovieQA

[Tapaswi et al., *CVPR* 2016]

Q) How does Patrick start winning Kat over?
A) By **knowing Kat's likes and dislikes**

LSMDC 16

[Rohrbach et al., *IJCV* 2017]



Q) Richard Parker _____ from the boat.
A) watches

- Solutions require information not available in visual content.
  eg) context, sound, script

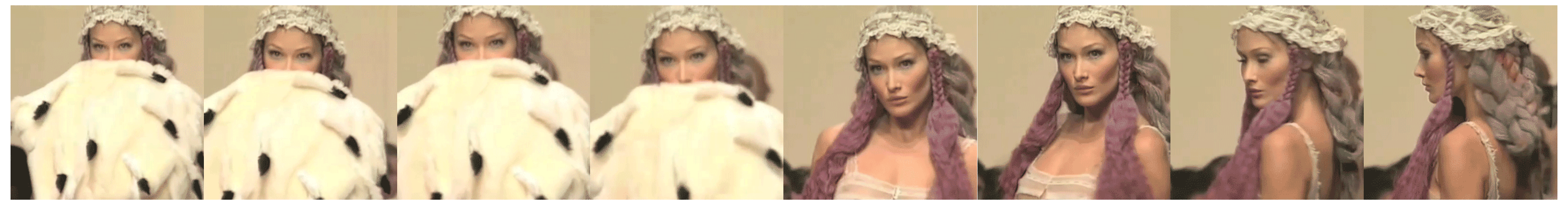
Can we define new tasks specifically for video VQA?

New Tasks for Video VQA

1. Counting Repetitions

Q) How many times does the animal pump arms? **A)** 2 times

2. Reasoning State Transitions

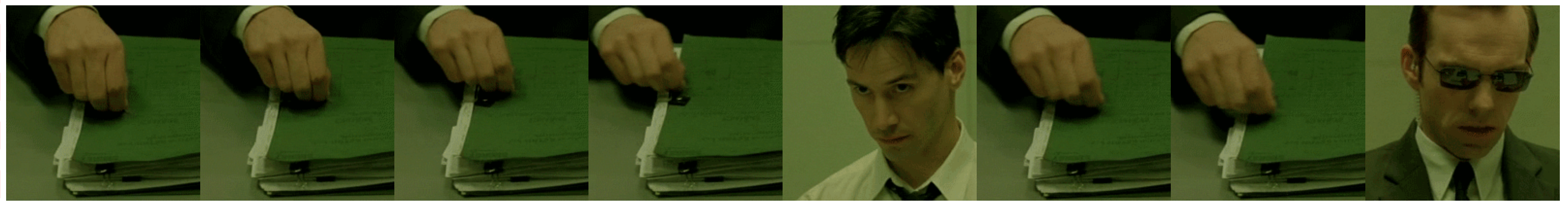


Q) What does the woman do after lowering the coat? **A)** Pivot around

A New Dataset for Video VQA: 165K QA Pairs

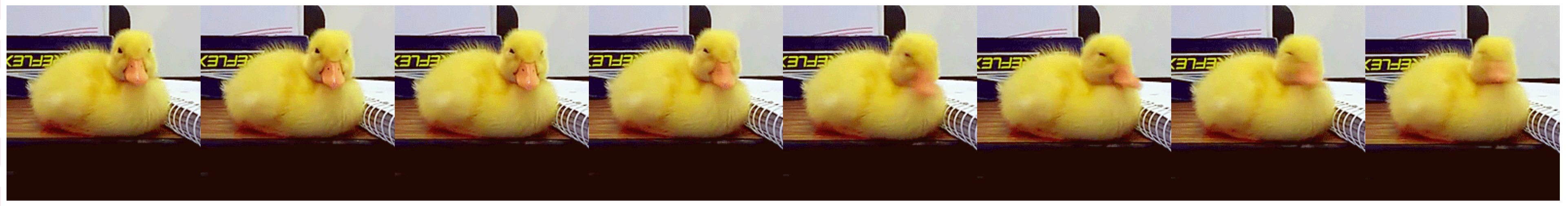
New Tasks for Video QA

a) Repetition Count: 30K



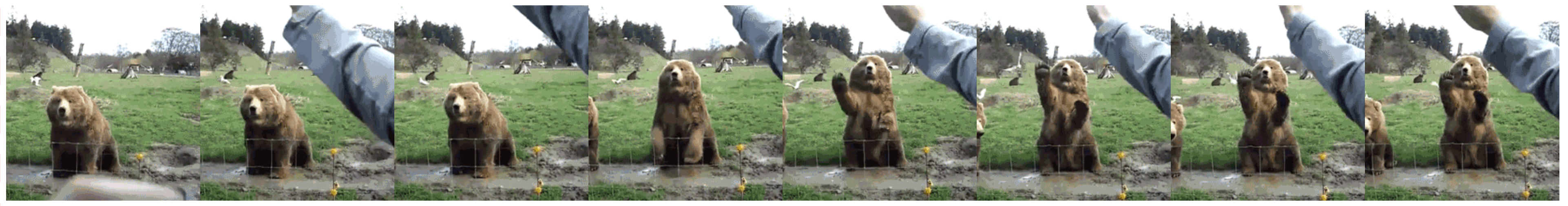
Q) How many times does the man wrap string? **A)** 5 times

b) Repeating Action: 23K



Q) What does the duck do 3 times? **A)** Shake head

c) State Transition: 59K



Q) What does the bear on right do after sitting? **A)** Stand

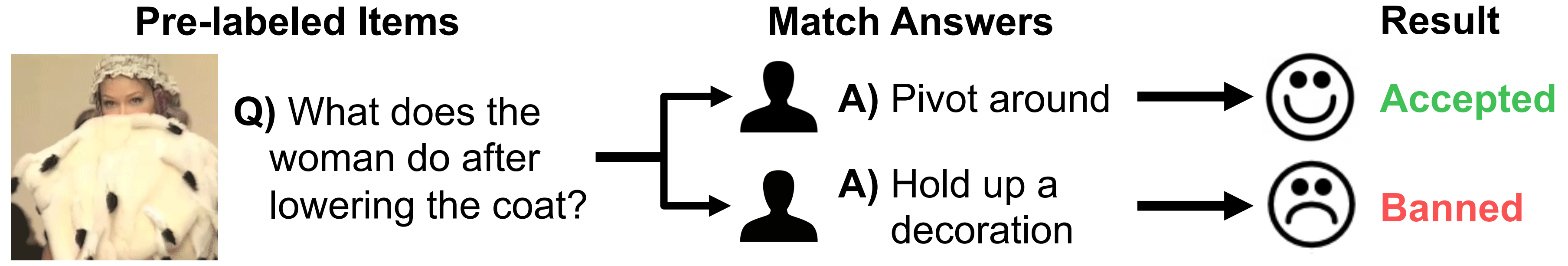
d) Frame QA: 53K

Q) What is dancing in the cup? **A)** Tree

Methods for Generating QA Pairs

a-c) Template-based. Crowdsourced via amazon mechanical turk Artificial Artificial Intelligence

- Strict quality control: Blacklist workers based on pre-labeled items



Pre-labeled Items **Match Answers** **Result**

Q) What does the woman do after lowering the coat?

A) Pivot around → Accepted

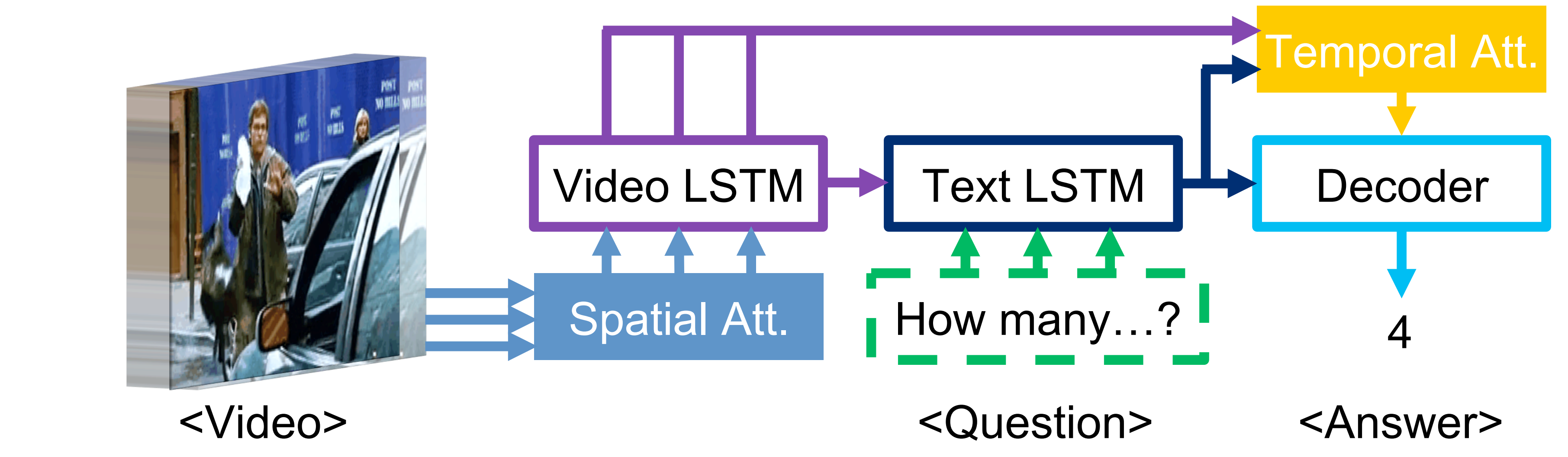
A) Hold up a decoration → Banned

- Synonyms are considered as correct answers
- Generate four wrong answers based on a cosine similarity of the verbs

d) NLP-based QA generation[1] using descriptions from TGIF dataset[2]

- Convert a declarative sentence to an interrogative sentence

A Novel Model for Video VQA: ST-VQA

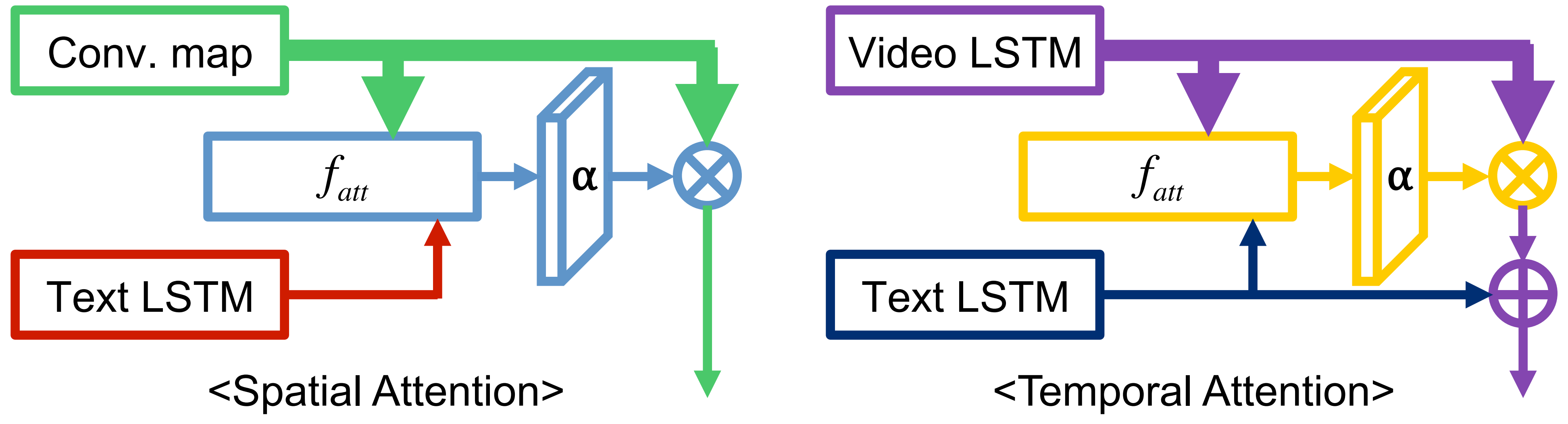


Video LSTM → Text LSTM → Decoder; Temporal Att.; Spatial Att.

How many…? 4

<Video> <Question> <Answer>

- Spatial[3] and temporal[4] attention modules



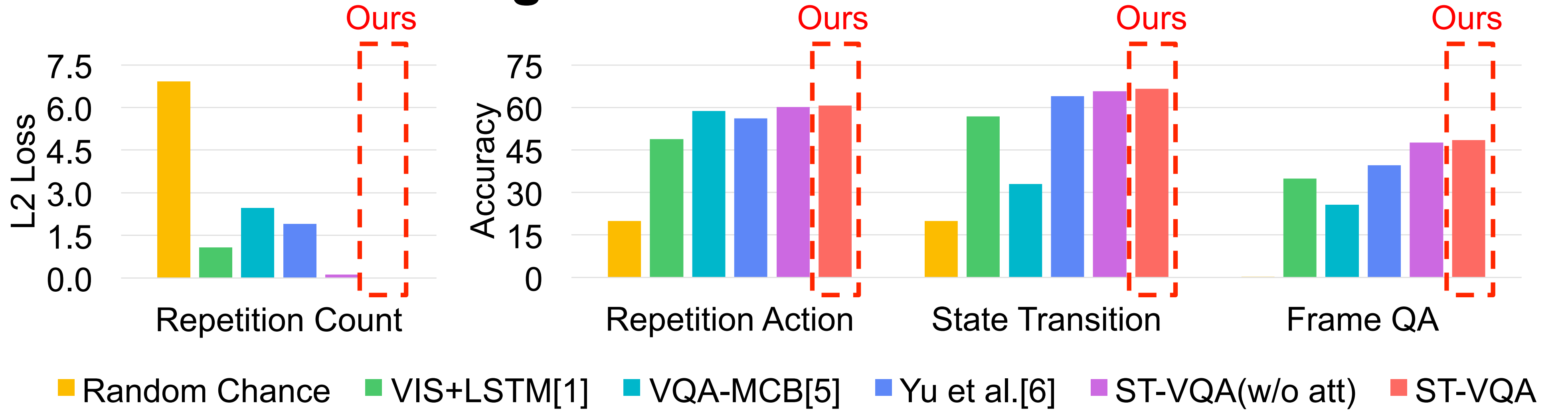
Conv. map, f_{att}, α, Text LSTM

<Spatial Attention>

Video LSTM, f_{att}, α, Text LSTM

<Temporal Attention>

Results and Findings



- **Video-based model** works better than image-based models
- Our model with **ST-attention** shows the best result.

Solving our tasks require **spatio-temporal reasoning**

References

[1] Ren et al., *Exploring Models and Data for Image Question Answering*, in NIPS 2015
[2] Li et al., *TGIF: A New Dataset and Benchmark on Animated GIF Description*, in CVPR 2016
[3] Xu et al, *Show, Attend and Tell: Neural Image Caption Generation with Visual Attention*, in ICML 2015
[4] Bahdanau et al, *Neural Machine Translation by Jointly Learning to Align and Translate*, in ICLR 2015
[5] Fukui et al., *Multimodal Compact Bilinear Pooling for Visual Question Answering and Visual Grounding*, in EMNLP 2016
[6] Yu et al., *End-to-end Concept Word Detection for Video Captioning, Retrieval, and Question Answering,* in CVPR 2017