Video Understanding

4 High-level Computer Vision for Video data

Video Understanding
High-level Computer Vision for Video data

Hudson and Manning. GQA: A New Dataset for Real-World Visual Reasoning and Compositional Question Answering. CVPR

GQA

Figure 1: Examples from the new GQA dataset for visual reasoning and compositional question answering:
*Is the **bowl** to the right of the **green apple**?*
*What type of **fruit** in the image is **round**?*
*What color is the **fruit** on the right side, red or **green**?*
*Is there any **milk** in the **bowl** to the left of the **apple**?*

Video Understanding
High-level Computer Vision for Video data

Hudson and Manning. GQA: A New Dataset for Real-World Visual Reasoning and Compositional Question Answering. CVPR

GQA

Pattern: What|Which <type> [do you think] <is> <dobject>, <attr> or <decoy>?
Program: **Select:** <dobject> → **Choose** <type>: <attr>|<decoy>
Reference: The food on the red object left of the small girl that is holding a hamburger
Decoy: brown

What color is the food on the red object left of the small girl that is holding a hamburger, yellow or brown?

Select: hamburger → **Relate:** girl,holding → **Filter** size: small → **Relate:** object, left → **Filter** color: red → **Relate:** food,on → **Choose** color: yellow | brown

Graph Normalization	Question Generation	Sampling and Balancing	Entailments Relations	New Metrics
• Ontology construction • Edge Pruning • Object Augmentation • Global Properties	• Patterns Collection • Compositional References • Decoys Selection • Probabilistic Generation	• Distribution Balancing • Type-Based Sampling • Deduplication	• Functional Programs • Entailment Relations • Recursive Reachability	• Consistency • Validity & Plausibility • Distribution • Grounding

Video Understanding
High-level Computer Vision for Video data

M. Grunde-McLaughlin et al. AGQA: A Benchmark for Compositional Spatio-Temporal Reasoning. CVPR

AGQA

Example compositional spatio-temporal questions:

Q: What did the person hold after putting a phone somewhere? A: bottle
Q: Were they taking a picture or holding a bottle for longer? A: holding a bottle
Q: Did they take a picture before or after they did the longest action? A: before

Generalization to novel compositions:

Q: Did the person twist the bottle after taking a picture? A: yes

Generalization to indirect references:

Q: Did the person twist the bottle? A: yes
Q: Did the person twist the object they were holding last? A: yes

Generalization to more compositional steps:

Q: What did they touch last before holding the bottle and after taking a picture, a phone or a bottle ? A: phone

Legend: objects relationships actions time

Video Understanding
High-level Computer Vision for Video data

M. Grunde-McLaughlin et al. AGQA: A Benchmark for Compositional Spatio-Temporal Reasoning. CVPR

AGQA

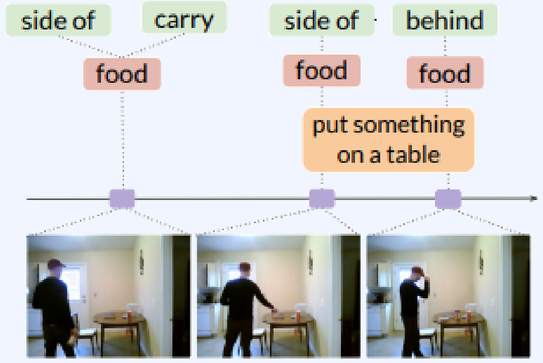
Video Understanding
High-level Computer Vision for Video data

M. Grunde-McLaughlin et al. AGQA: A Benchmark for Compositional Spatio-Temporal Reasoning. CVPR

AGQA

Input

Spatio-temporal scene graph

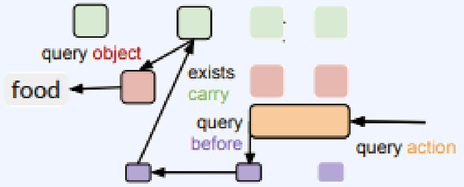


Video

Our benchmark generation process

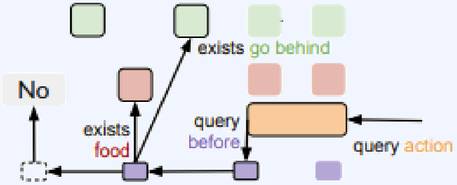
Template #1: What did they <relationship><time><action>?

Program #1:
```
for frame before put something on a table:
  if exists(carry):
    return query(object)
```



Template #2: Did they <relationship> <object> <time><action>?

Program #2:
```
for frame before put something on a table:
  if exists(go behind-food):
    return Yes
return No
```



Output

Question answer #1:
What did they carry before putting something on a table? - food

Indirect action reference for #1:
What did they carry before the shortest action? - food

Question answer #2:
Did they go behind some food before putting something on a table? - No

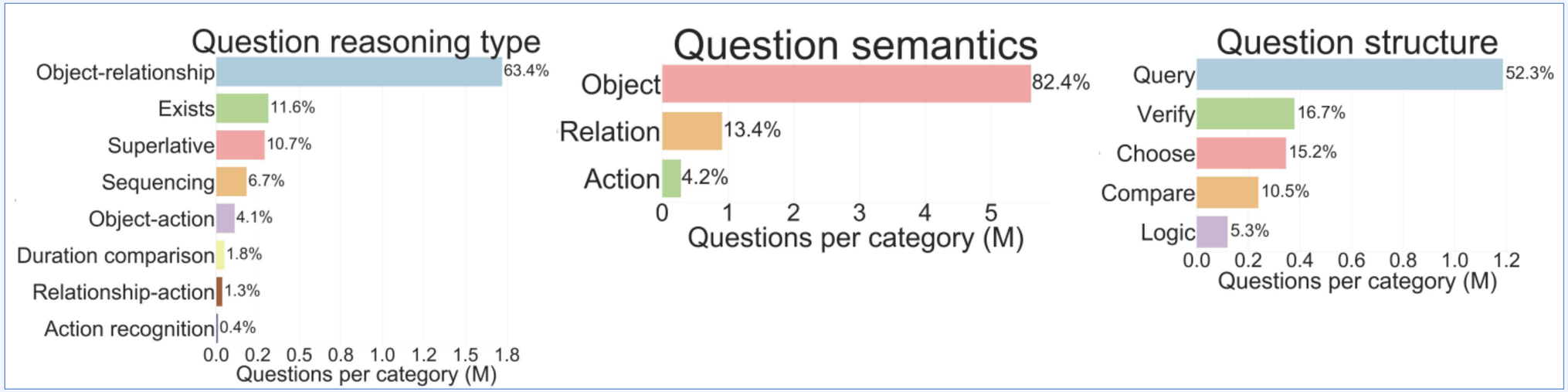
Indirect object reference for #2:
Did they go behind the object they carried before putting something on a table? - No

Legend: objects relationships actions time

Video Understanding
High-level Computer Vision for Video data

M. Grunde-McLaughlin et al. AGQA 2.0: An Updated Benchmark for Compositional Spatio-Temporal Reasoning.

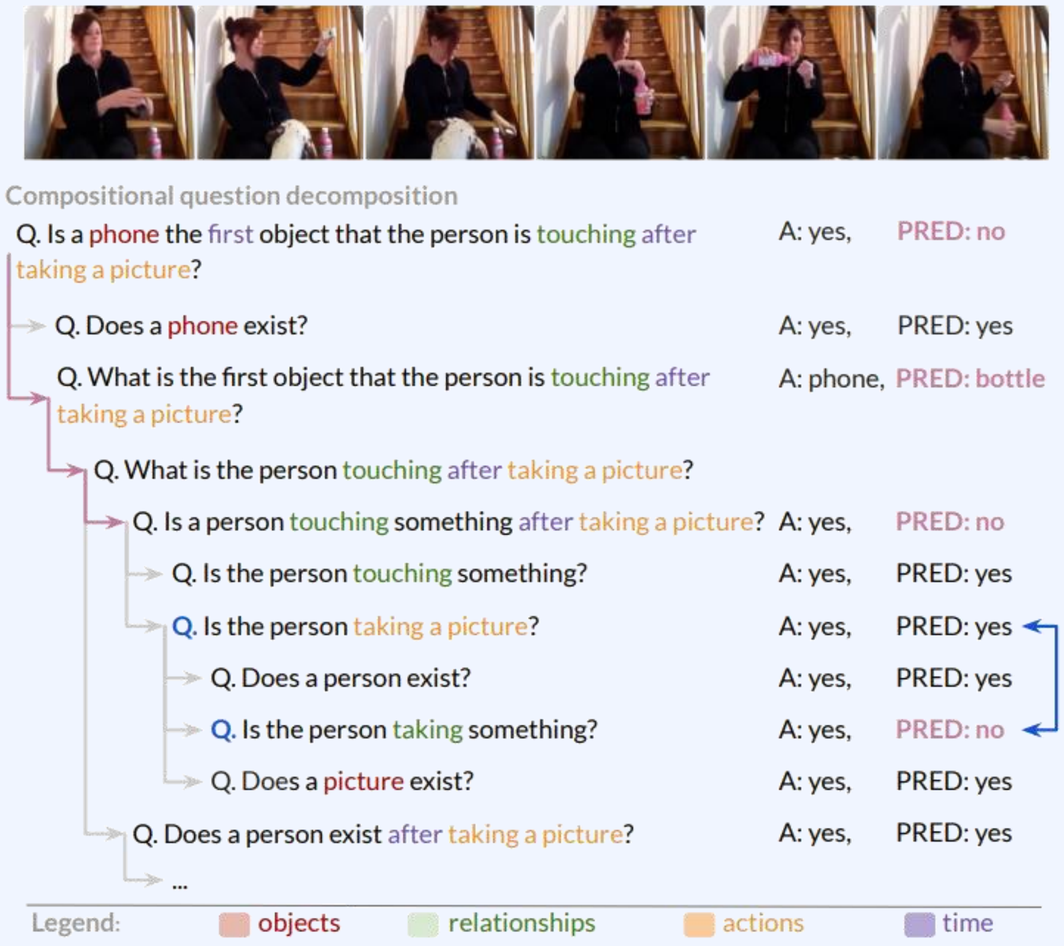
AGQA



Video Understanding
High-level Computer Vision for Video data

M. Gandhi et al. Measuring Compositional Consistency for Video Question Answering. CVPR

AGQA-Decomp



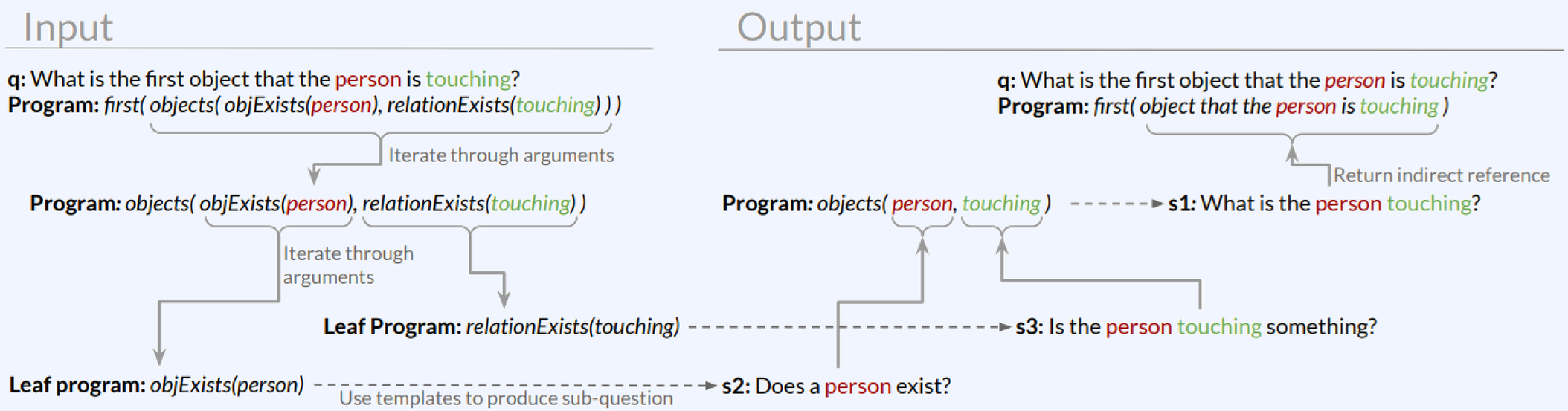
M. Gandhi et al. Measuring Compositional Consistency for Video Question Answering. CVPR

AGQA-Decomp

Sub-question type	Description	Example
`Object exists`	To verify if an object exists	Does a doorway exist?
`Relation exists`	To verify if a relationship exists	Is the person holding something?
`Interaction`	To verify if there is a particular relationship between person and an object	Is the person touching a dish?
`Interaction temporal loc.`	A filter on an interaction type question	Is the person holding a book while smiling at something?
`Exists temporal loc.`	A condition on object/relationship exists question	Does a phone exist after looking in the mirror?
`First/last`	Getting the first/last instance of the given object	What is the first object that the person is above before walking through the doorway?
`Longest shortest action`	Getting the longest/shortest action	What does the person do for the longest amount of time?
`Conjunction`	Get a new exists question by combining two interaction questions with a conjunction	Is the person in front of the mirror and behind the table while looking in the mirror?
`Choose`	Compares between two objects, actions, relationships, or time lengths	Is the doorknob or the dish the first object that the person is holding?
`Equals`	Compares two objects and verifies if they are the same Verifies if the given action is longer/shorter than the other one	Is the doorway the object they are interacting with while holding a dish?

M. Gandhi et al. Measuring Compositional Consistency for Video Question Answering. CVPR

AGQA-Decomp



M. Gandhi et al. Measuring Compositional Consistency for Video Question Answering. CVPR

AGQA-Decomp

Composition rules	Description	Example
`Interaction`	Verify if an interaction exists	*q* : Is a person holding a doorway? *s1*: Does a person exist? *s2*: Is a person holding something? *s3*: Does a doorway exist?
`Temporal loc.` `(After, before, while, between)`	Combine two interaction or exists questions using a temporal localizer	*q*: Is the person touching a doorway before smiling at something? *s1*: Is the person touching a doorway? *s2*: Is a person smiling at something?
`First/last`	Getting the first/last occurrence from a set of object/actions	*q*: What is the first object that the person is holding? *s1*: What is the person holding?
`Conjunction` `(And, xor)`	Combine two interaction questions using a conjunction	*q*: Is the person putting some clothes and behind a book before walking through the doorway? *s1*: Is the person putting some clothes before walking through the doorway? *s2*: Is the person behind a book before walking through the doorway?

`Choose`
`(Choose (object/Time)`
`longer/shorter choose)`
	Chooses one of two possible options	*q*: Is the doorway or the book the first object they were in front of?
		s1: Is the doorway the first object they were in front of?
		s2: Is the book the first object they were in front of?

`Equals`	Compares two objects/actions to verify if they are the same	*q*: Is a book the first object that the person is carrying? *s1*: Does a book exist? *s2*: What is the first object that the person is carrying?

Video Understanding
High-level Computer Vision for Video data

M. Gandhi et al. Measuring Compositional Consistency for Video Question Answering. CVPR

AGQA-Decomp

Compositional accuracy

If the model answers all children questions correctly, does it answer the parent question correctly?

a question that is answered correctly
a question that is answered incorrectly

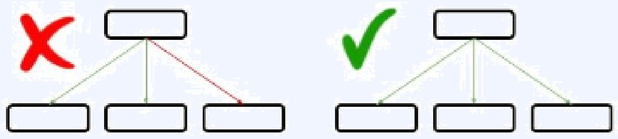
Right for the wrong reasons

If the model answers a child question incorrectly, does it still answer the parent question correctly?

a question that is answered correctly
a question that is answered incorrectly

Internal Consistency

Do the model's answers reflect a consistent understanding of visual events?

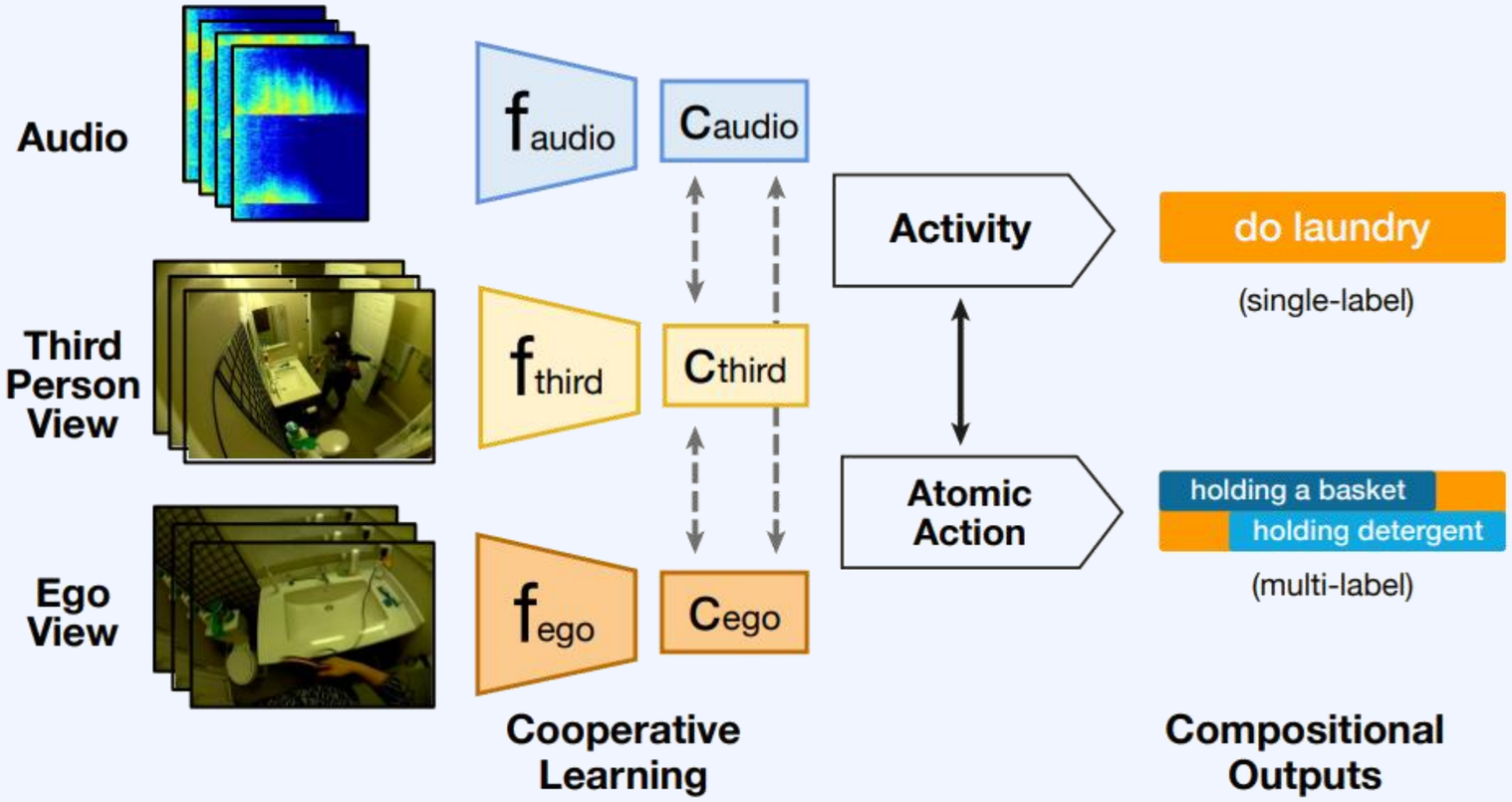


parent and child answers are consistent
parent and child answers are inconsistent

Video Understanding
High-level Computer Vision for Video data

N. Rai et al . Home Action Genome: Cooperative Compositional Action Understanding. CVPR

HOMAGE



Video Understanding
High-level Computer Vision for Video data

N. Rai et al . Home Action Genome: Cooperative Compositional Action Understanding. CVPR

HOMAGE

Video Understanding
High-level Computer Vision for Video data

N. Rai et al . Home Action Genome: Cooperative Compositional Action Understanding. CVPR

HOMAGE