Multimodal CL and NLP: Combining Language and Vision for (Computational) Semantics

Carina Silberer

IMS, University of Stuttgart

CL Fall School 2024, Passau
16 - 20 September 2024

Outline

Multimodality

[Liang et al., 2022]

Characteristics

- Modality:
  - raw: speech, audio, images, gaze etc.
  - abstract: language, objects etc.
- Multimodal: Involvement of multiple modalities
- Multiple modalities are *heterogeneous*, *connected* and *interacting*

Multimodality

Relevance

- Humans seamlessly integrate diverse sensory inputs to act with / interact in the world
- Machines that are capable of human-like cognitive functions must similarly be able to process multimodal data
- Provides machines with contextual awareness and the ability to adapt to real-world scenarios

Introduction: Multimodal NLP

- Deals with problems at the intersection of language and vision ("Vision & Language")
- Focuses on language understanding with computer vision (and audio processing etc.)
- Our focus: Combined image and language understanding
- Includes also speech and other modalities (not topic in this course)

References

Experience Grounds Language [Bisk et al., 2020]
Five World Scopes: Corpus (our past), Internet, **Perception**, Embodiment, Social

History of Semantics Grounded in Perception and Action

Publications on Multimodality and Grounding in ACL+EMNLP

year	# papers	track
2012	2	–
2019	49	Speech, Vision, Robotics, Multimodal and Grounding
2020	64	Language Grounding to Vision, Robotics and Beyond

NAACL 2021: 3 different workshops related to multimodality[1]

[1] https://2021.naacl.org/program/workshops/

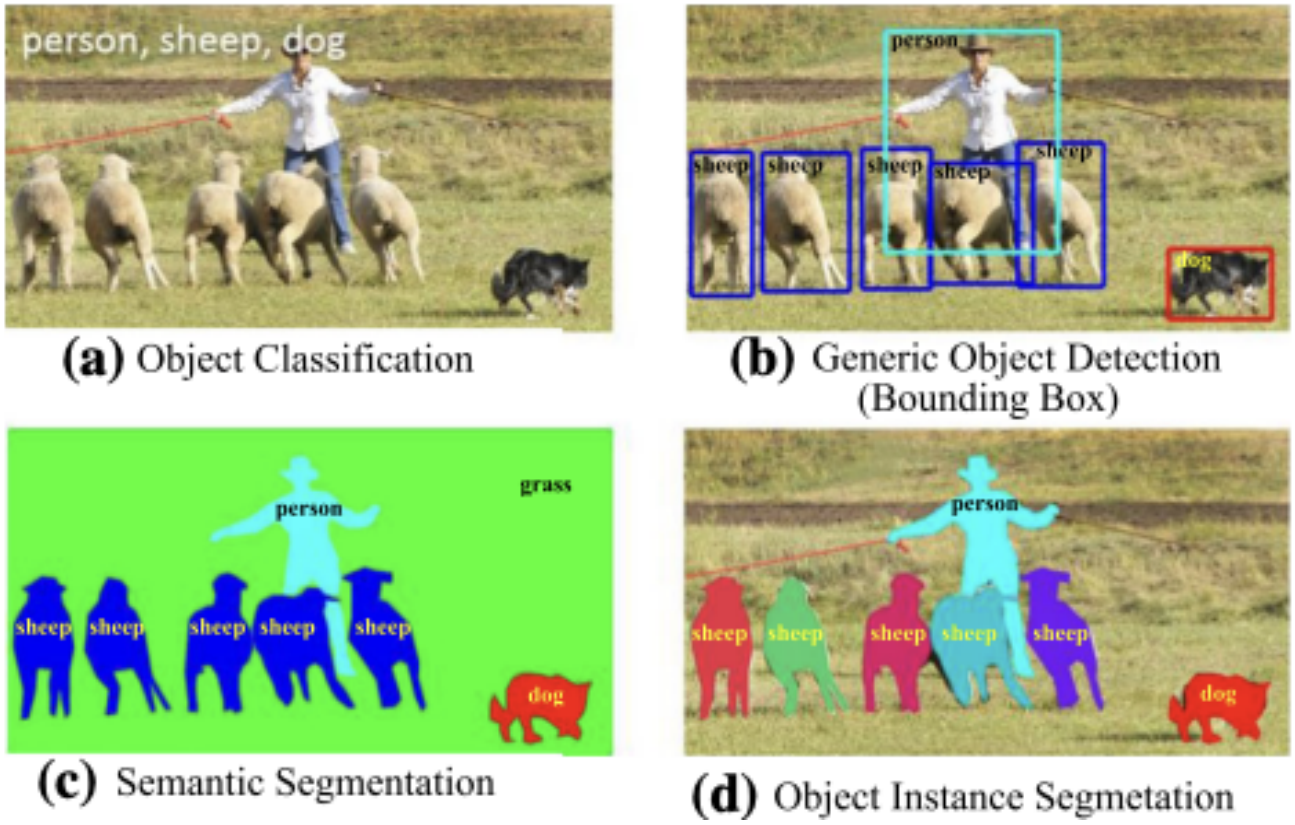
Language & Vision

Connected research fields and topics

- Computer Vision (CV): object recognition (image classification and object detection), action and attribute recognition, ...
  Role of language: leverage language to achieve a deeper image understanding, or as source of supervision
- Machine Learning and Pattern Recognition: deep neural networks – representation learning, classification, generation, ...
- Natural Language Processing/Computational Linguistics: language generation, language comprehension, ...
  Role of vision: obtain richer (more complete) semantic knowledge; situated language understanding

Outline

Computer Vision Tasks: Object Recognition



(a) Object Classification

(b) Generic Object Detection (Bounding Box)

(c) Semantic Segmentation

(d) Object Instance Segmetation

source: *[Liu et al., 2020]*

(a) Object Classification is also called *image classification*

Computer Vision is also known as
Image Analysis, *Scene Analysis*, *Image Understanding*

Language & Vision: Machine Learning

(Deep) Neural Architectures

- Computer Vision: convolutional neural networks (CNNs), visual transformers, latent diffusion models
- NLP: recurrent neural networks (RNNs), LSTMs, transformers
- autoencoders, attention mechanism

Language & Vision: Machine Learning

Methods and Representations

- How de we meaningfully represent images or image regions (e.g., objects)?
- How do we represent the meaning of textual units (e.g., words, phrases, sentences)

$\Rightarrow$ Fixed-length, real-valued vectors

- How do we obtain / learn these representations?

Language & Vision: Machine Learning

Computer Vision Methods and Visual Representations

(Deep) Neural Architectures for Computer Vision

- Image classification: AlexNet [19], VGGNet [20], GoogLeNet [22], and ResNet [21], Inception, ... (CNNs); Vision Transformer (ViT)
- Object detection: region-based CNN (R-CNN) [23], Fast R-CNN [24], and Faster R-CNN [25], YOLOS, ViT [Dosovitskiy et al., 2021] ...

$\Rightarrow$ **Extract feature vectors from hidden layers which represent the input image, or image regions of it (e.g., objects)**

- Image generation: diffussion models; GAN (generative adversarial networks)

NLP: Learning Language Representations

- Distributional Hypothesis: Words that appear in similar contexts tend to be similar in meaning (Harris, 1954)

$\Rightarrow$ Learn meaning of textual units from text:
through lexical co-occurrence within and beyond sentences
(through sentence and document context)

- *Embedding*: a mapping from a one-hot vector representing a textual unit (word, subword, phrase, ...) to a fixed-length vector of real-valued numbers

$\Rightarrow$ Ought to represent the meaning of linguistic units

Language Meaning

The meaning of ***sheep*** ...

Generic, Conceptual, Language-based

> Most *sheep grow woolly coats*. [...] Mostly *sheep eat grass, legumes, forbs*, and other pasture *plants*. They especially *love forbs*. In fact, it is usually their first choice of food in a pasture. A forb is a broad-leaf plant other than grass. ...

Language & Vision: Machine Learning

NLP Methods and Language Representations (embeddings)

(Deep) Neural Architectures for NLP

- Neural network language model (NNLM): estimates the probability of a word sequence
- Continuous bag-of-words model, skip-grams, and global vectors (GloVe): captures co-occurrence patterns of words
- Latest generation models: Contextualised embeddings (e.g., ELMo, BERT, RoBERTa, GPT*, Gemini, LLAMA, etc.)

Language & Vision

Difference of CV and NLP

- CV: usually always situated, but not necessarily instance-based
  But: simplified view on language
- NLP: global (conceptual/static)
  or situated in specific context (contextualised) $\rightarrow$ dynamic;
  – usually not situated (embedded) in physical environment
  – no use of visuals / computer vision
  But: You can't learn language from the radio or internet

Language Meaning

The meaning of ***sheep*** ...

Generic, Conceptual, Language-based

> Most *sheep grow woolly coats*. [...] Mostly *sheep eat grass, legumes, forbs*, and other pasture *plants*. They especially *love forbs*. In fact, it is usually their first choice of food in a pasture. A forb is a broad-leaf plant other than grass. ...

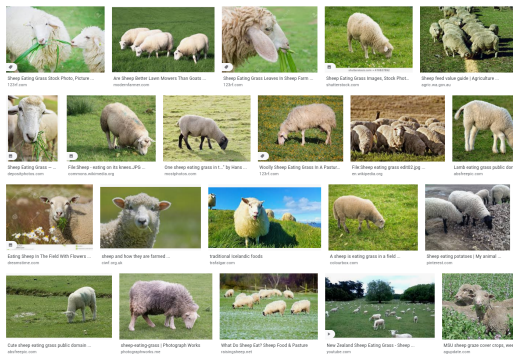
Language Meaning

The meaning of ***sheep*** ...

Generic, Conceptual, Language-based

Most *sheep grow woolly coats*. [...] Mostly *sheep eat grass, legumes, forbs*, and other pasture *plants*. They especially *love forbs*. In fact, it is usually their first choice of food in a pasture. A forb is a broad-leaf plant other than grass. ...

Generic, Grounded in Visual World



+

Most *sheep grow woolly coats*. [...] Mostly *sheep eat grass, legumes, forbs*, and other pasture *plants*. They especially *love forbs*. In fact, it is usually their first choice of food in a pasture. A forb is a broad-leaf plant other than grass. ...

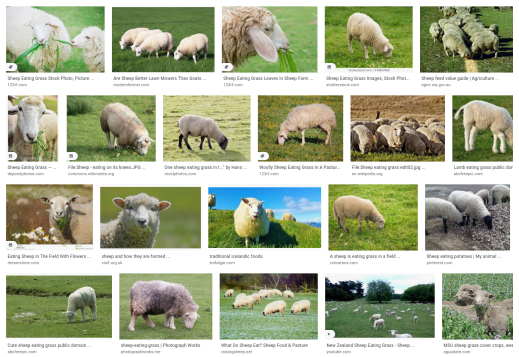
Language Meaning

The meaning of ***sheep*** ...

Generic, Conceptual, Language-based

Most *sheep grow woolly coats*. [...] Mostly *sheep eat grass, legumes, forbs*, and other pasture *plants*. They especially *love forbs*. In fact, it is usually their first choice of food in a pasture. A forb is a broad-leaf plant other than grass. ...

Generic, Grounded in Visual World



+

Most *sheep grow woolly coats*. [...] Mostly *sheep eat grass, legumes, forbs*, and other pasture *plants*. They especially *love forbs*. In fact, it is usually their first choice of food in a pasture. A forb is a broad-leaf plant other than grass. ...

Contextual, Situated in Visual World

Bob the *sheep* is eating a *muffin*.

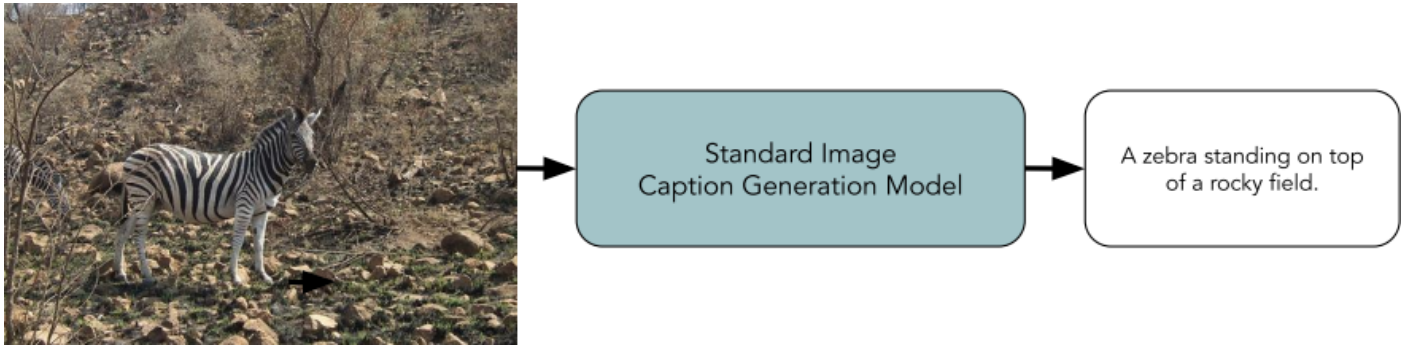
Outline

Language & Vision Tasks

Which "classic" problems
connected to visual and language understanding
has been L&V research seeking to solve?

Typical and highly researched topics

Image Captioning/Description Generation



source: *[Mogadala et al., 2021]*

Ground Truth Caption: Two brown bears playing in a field together.

Generated Caption: Two brown bears playing on top of a lush green field.

Ground Truth Caption: A plate of breakfast food with a silver tea pot.

Generated Caption: A close up of a plate of food with a folk and a knife on a table.

Fig. 11. Captions generated by Wu et al. [149] on some sample images from the MS COCO dataset.

source: *[Hossain et al., 2018]*

Image Captioning/Description Generation

Underlying Unimodal Tasks and Challenges

- Natural Language Generation, log-likelihood language modeling

Why Relevant?

- Computer Vision: Tests image understanding
- Applications: Help for visually impaired people, alternative to image (mobile data)

Visual Question Answering

Question : Are the bristles turned upward?

Original Image | **no**

Complementary Image | **yes**

source: *VQA v2*

Demo: `http://vqa.cloudcv.org/`

Visual Question Answering

Who is wearing glasses?

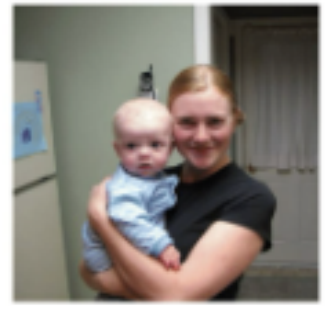
man

woman

Where is the child sitting?

fridge

arms



Is the umbrella upside down?

yes

no

How many children are in the bed?

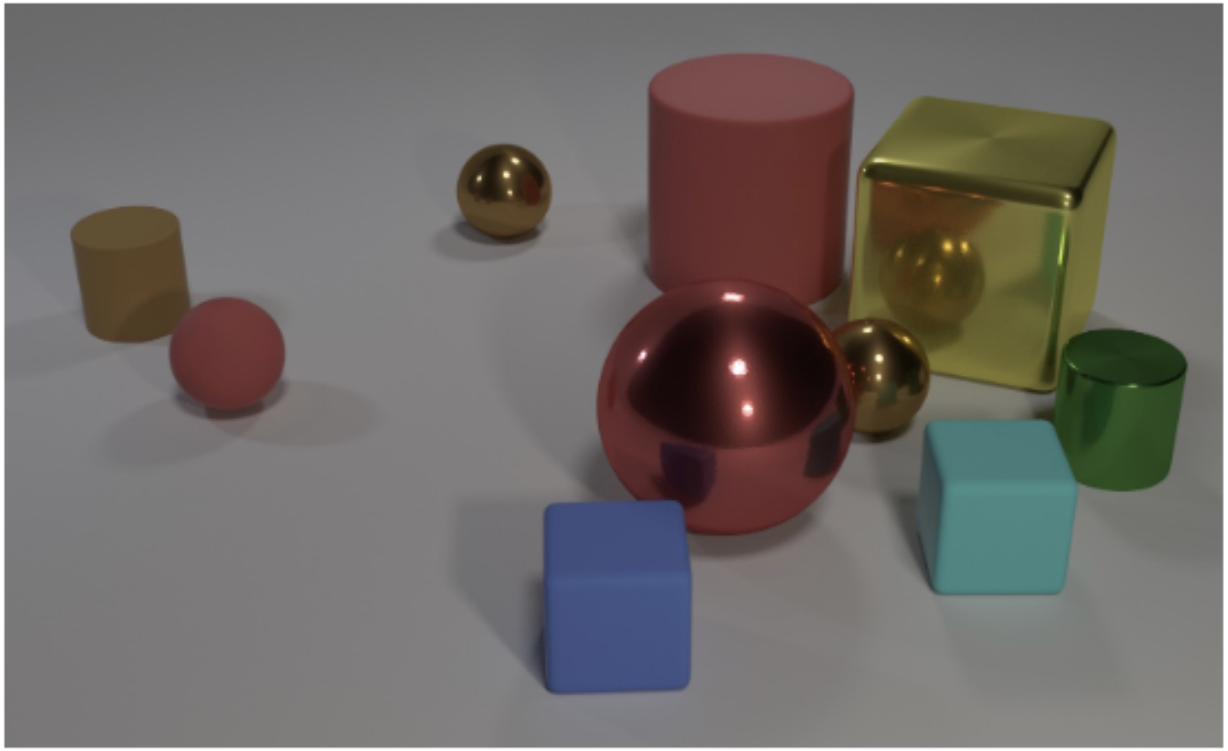
2

1

https://visualqa.org/

Visual Question Answering

Questions in CLEVR test various aspects of visual reasoning including **attribute identification**, **counting**, **comparison**, **spatial relationships**, and **logical operations**.



Q: Are there an **equal number** of **large things** and **metal spheres**?

Q: **What size** is the **cylinder** **that is left of** the **brown metal** thing **that is left of** the **big sphere**?

Q: There is a **sphere** with the **same size as** the **metal cube**; is it **made of the same material as** the **small red sphere**?

Q: **How many** objects are **either** **small cylinders** or **red** things?

https://cs.stanford.edu/people/jcjohns/clevr

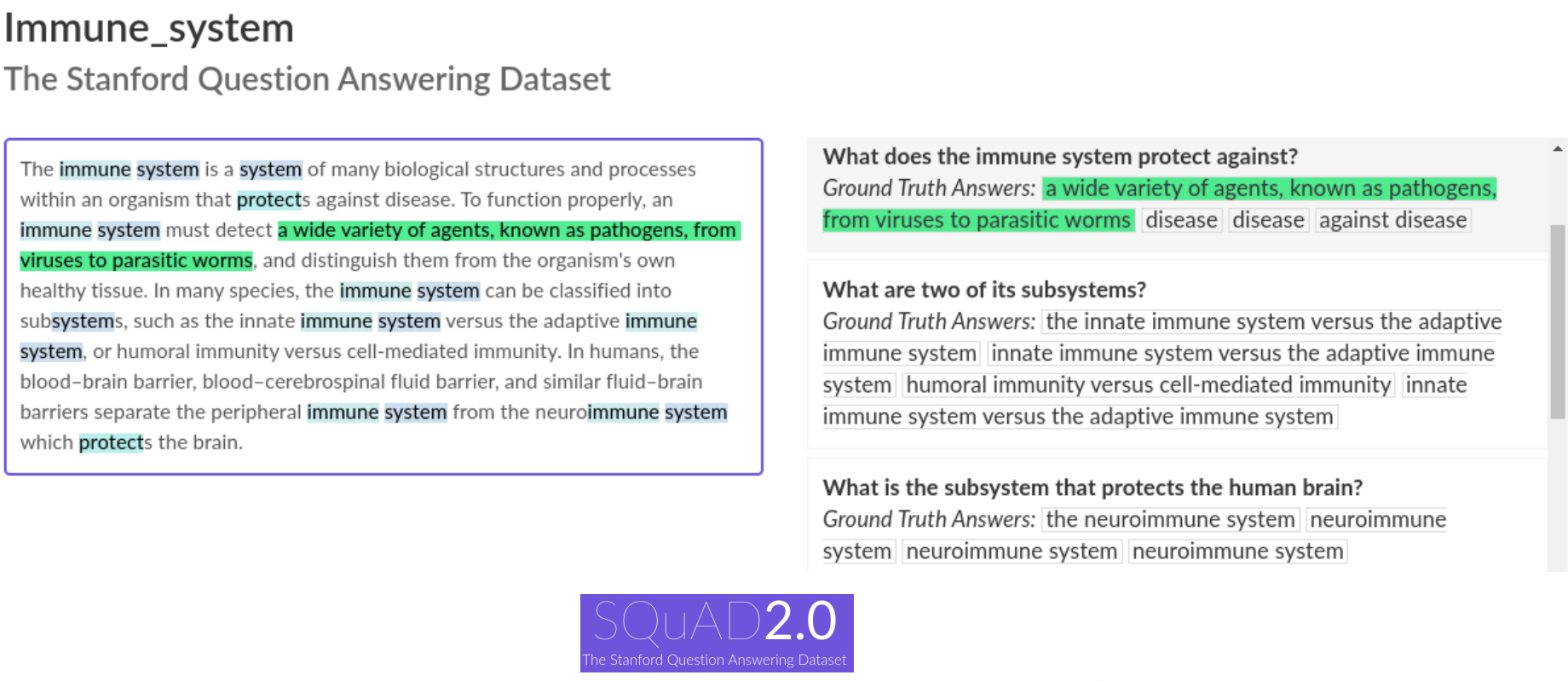
Visual Question Answering

https://rajpurkar.github.io/SQuAD-explorer/

Underlying Unimodal Tasks and Challenges

- Language-based Question Answering; on/in text, dialogue and knowledge bases

Goal language comprehension



source: *[Rajpurkar & Jia,'18]*

Visual Question Answering

Underlying Unimodal Tasks and Challenges

- Language-based Question Answering; on/in text, dialogue and knowledge bases

Why Relevant?

- Computer Vision: Tests visual comprehension and reasoning

?? Multimodal understanding & reasoning (language + vision) e.g., `https://hucvl.github.io/recipeqa/`

- Applications: Help for visually impaired people; information retrieval

Outline

Multimodal (Bimodal) Representation Learning

NLP: Learning Language Representations

Multimodal (Bimodal) Representation Learning

NLP: Learning Language Representations

- Distributional Hypothesis: Words that appear in similar contexts tend to be similar in meaning (Harris, 1954)

$\Rightarrow$ Learn meaning of textual units from text:
through lexical co-occurrence within and beyond sentences
(through sentence and document context)

Multimodal (Bimodal) Representation Learning

NLP: Learning Language Representations

- Distributional Hypothesis: Words that appear in similar contexts tend to be similar in meaning (Harris, 1954)

$\Rightarrow$ Learn meaning of textual units from text:
through lexical co-occurrence within and beyond sentences
(through sentence and document context)

- How can we learn aspects of meaning connected to the physical world, and situated in the world?

References

Bisk et al. (2020): Experience Grounds Language
Five World Scopes: Corpus (our past), *Internet* (our present), **Perception**, Embodiment, Social

You cannot learn language by listening to the radio!

How is language connected to the world?

- Reference or Intentionality: relationship that holds between language use and the world
- The meaning of words and utterances depend on the physical environment (and social and psychological environment)
- Concepts (mental structures) are grounded (connected) to the world

Multimodal (Bimodal) Representation Learning

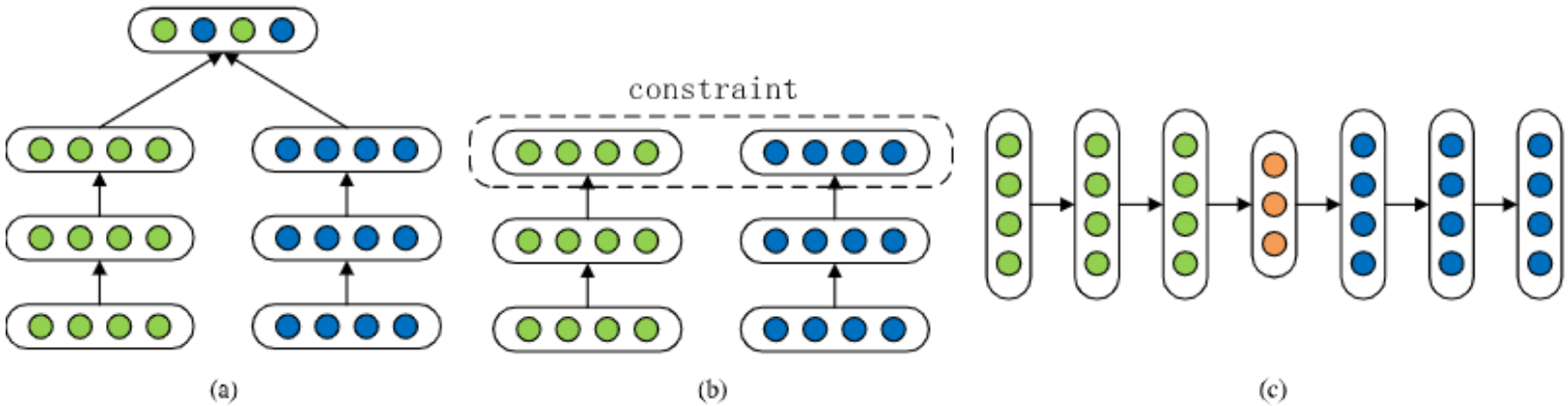


FIGURE 2. Three types of frameworks about deep multimodal representation. (a) Joint representation aims to learn a shared semantic subspace. (b) Coordinated representation framework learns separated but coordinated representations for each modality under some constraints. (c) Encoder-decoder framework translates one modality into another and keep their semantics consistent.

source: *[Guo et al.,'19]*

Multimodal (Bimodal) Representation Learning

[Liang et al., 2022]

Multimodal (Bimodal) Representation Learning

Underlying Unimodal Tasks and Challenges

- Representation learning for image data
- Representation learning for language data (text)

Goal Represent the content/meaning aspects of visuals and language

Why Relevant?

- Cognitive Science: How do humans represent meaning?
- NLP + CV: (General) representations for L&V methods, capturing shared and complementary information

?? How do we learn these, and fuse the modalities? How do we assess *what* they actually capture?

- Applications: Useful for all tasks in NLP, CV, L&V

Outline

Language Grounding: Referring Expression Comprehension

The *Steinpilzknödel*

www.veggielicious.de/rezepte/suedtiroler-knoedeltris

Language Grounding: Referring Expression Comprehension

The *Steinpilzknödel*

www.veggielicious.de/rezepte/suedtiroler-knoedeltris

Language Grounding: Referring Expression Comprehension

The Steinpilzknödel – ???

www.kuechengoetter.de/rezepte/abgeschmolzene-steinpilzknoedel-37162

Language Grounding: Referring Expression Comprehension

???

www.kuechengoetter.de/rezepte/abgeschmolzene-steinpilzknoedel-37162

Language Grounding: Referring Expression Comprehension

RefCOCO TestA

curly redhead

RefCOCO TestB

top most donut left side

RefCOCO+ TestA

guy in pink

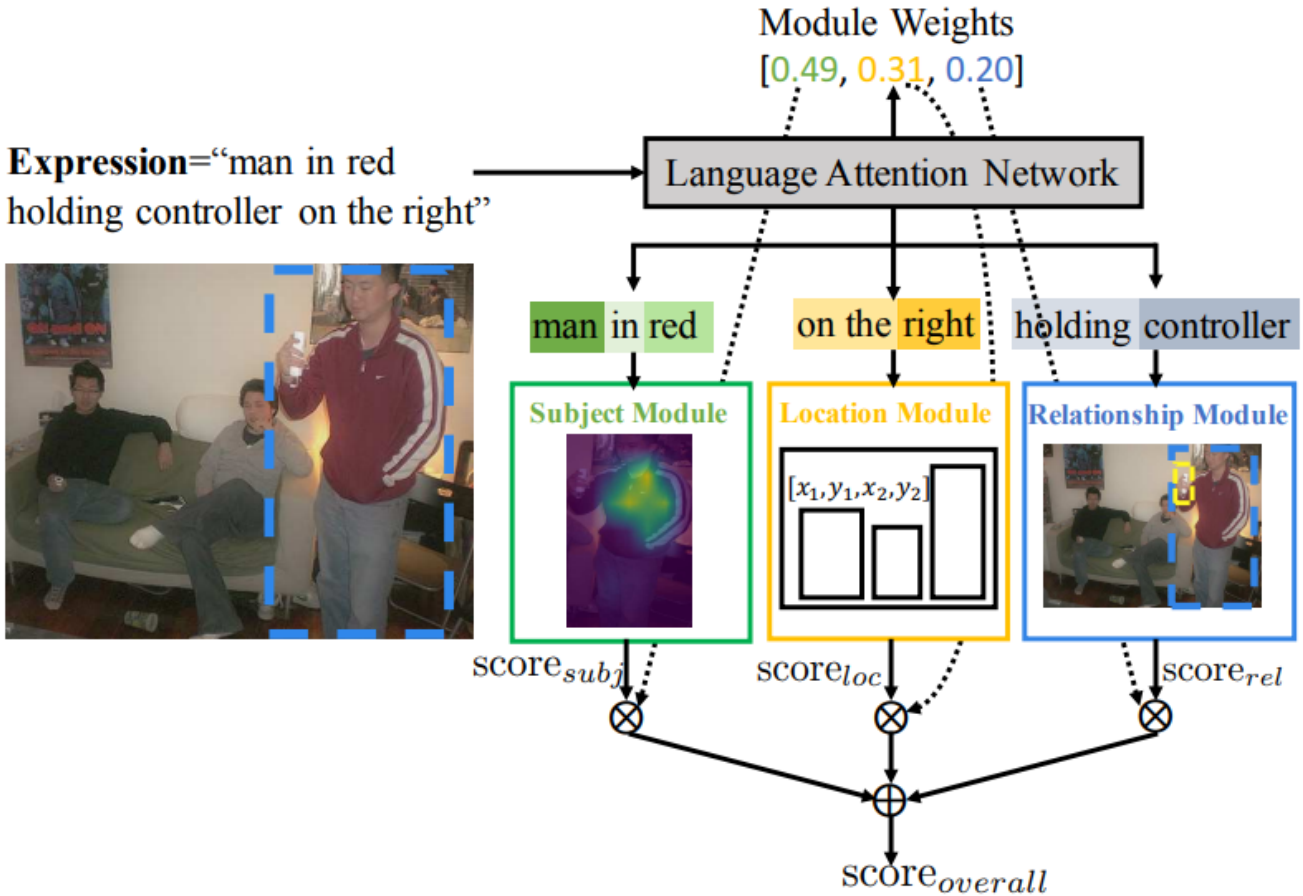
RefCOCO+ TestB

red meat sandwich

source: *Liu et al., 2017*

Language Grounding: Referring Expression Comprehension

Modular Attention Network (MAttNet)

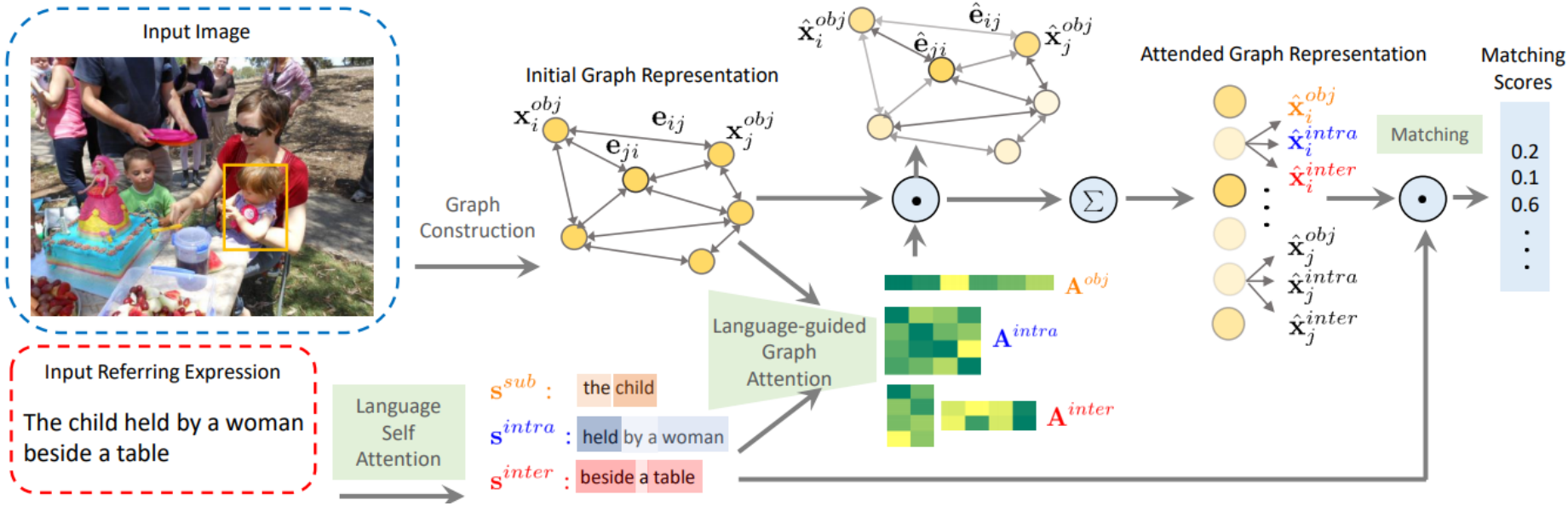


Yu et al.,'18

A given expression is attentionally parsed by a subject, a location, and a relationship module.

Language Grounding: Referring Expression Comprehension

Neighbourhood Watch: Referring Expression Comprehension via Language-guided Graph Attention Networks



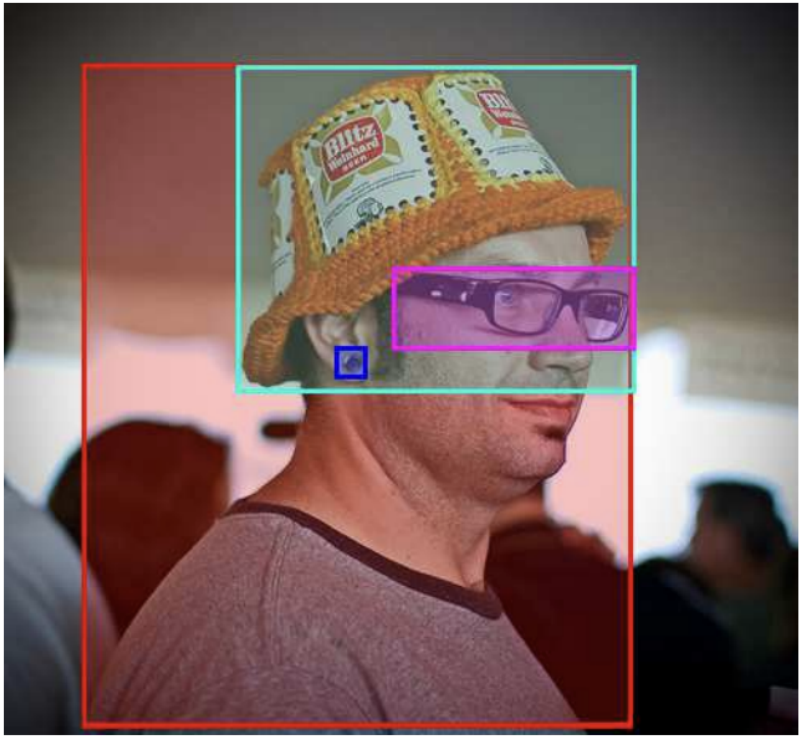
[Wang et al., 2018]

The network has three modules: language self-attention module, language-guided graph attention module, matching module

Language Grounding: Phrase Grounding

[Plummer et al., 2015, Nakayama et al., 2020]

Flickr30k Entities



A man with **pierced ears** is wearing **glasses** and **an orange hat**.
A man with **glasses** is wearing **a beer can crotched hat**.
A man with **gauges** and **glasses** is wearing **a Blitz hat**.
A man in **an orange hat** starring at **something**.
A man wears **an orange hat** and **glasses**.

- With coreference chains across captions

Language Grounding: Phrase Grounding

[Plummer et al., 2015, Nakayama et al., 2020]

Flickr30k Entities Japanese

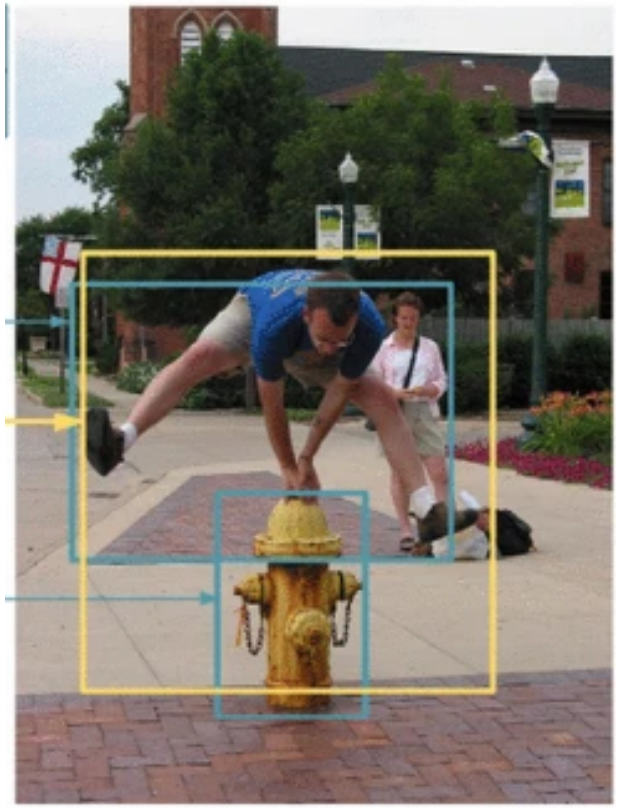
Flickr30k Entities (Plummer et al., 2017)	(1) An old woman wearing a yellow jacket and blue jeans trying to choose some vegetables from a street stand. (2) A gray-haired woman in a yellow jacket looks at vegetable produce at a farmer 's market stall.
Our Japanese Translations	(1) 露店から野菜を選ぼうとしている黄色いジャケットと青いジーンズを着ている年配の女性。 (2) 黄色いジャケットを着ている白髪の女性がファーマーズマーケットの売店で農産物を見ている。

- With coreference chains across captions
- With Japanese correspondence

Language Grounding: Phrase Grounding

[Krishna et al., 2017] huggingface.co/datasets/ranjaykrishna/visual_genome

Visual Genome

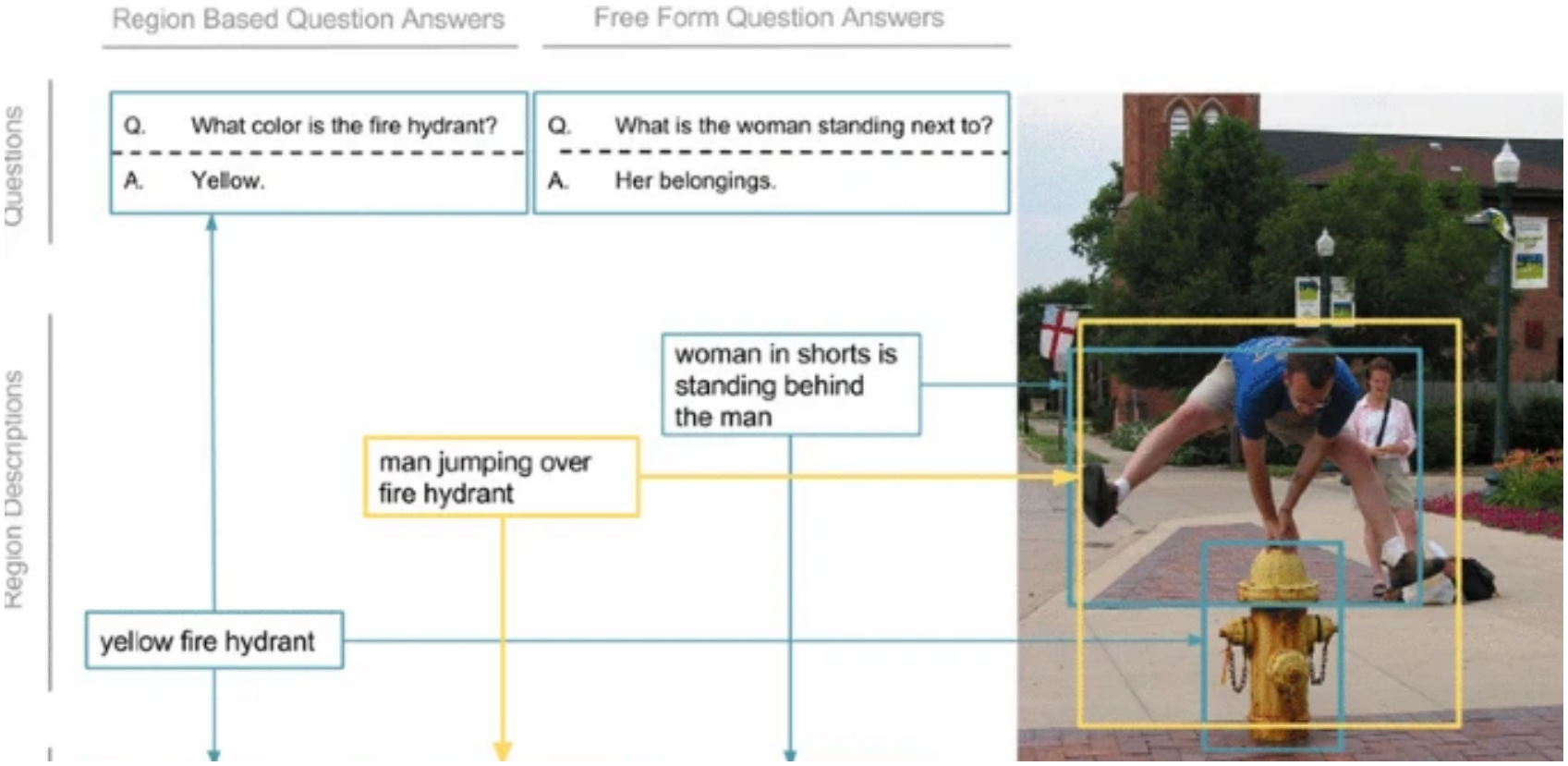


- Densely labeled images with attributes, nouns, relations, region descriptions

Language Grounding: Phrase Grounding

[Krishna et al., 2017] huggingface.co/datasets/ranjaykrishna/visual_genome

Visual Genome

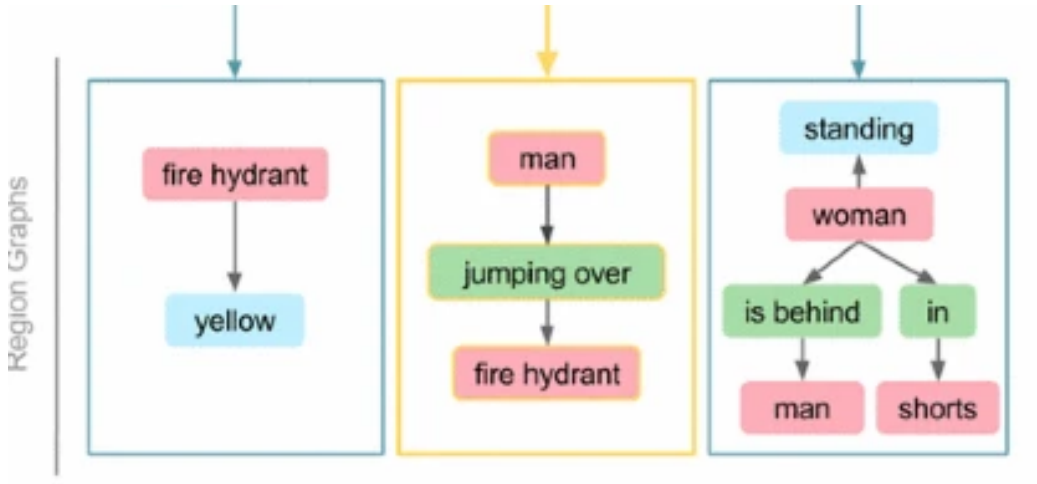


- For grounded question answering

Language Grounding: Phrase Grounding

[Krishna et al., 2017] `huggingface.co/datasets/ranjaykrishna/visual_genome`

Visual Genome

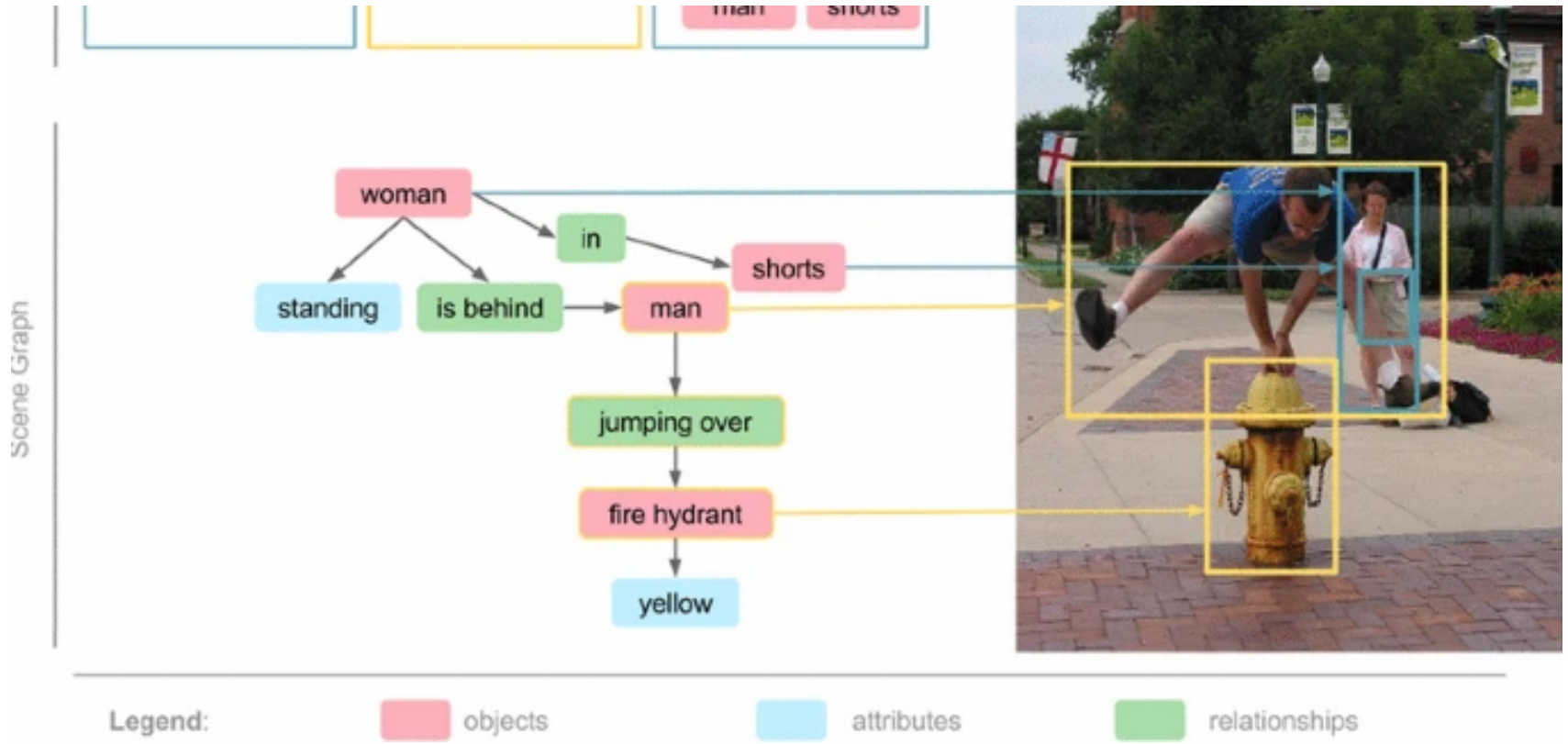


- Graph-based region descriptions

Language Grounding: Phrase Grounding

[Krishna et al., 2017] huggingface.co/datasets/ranjaykrishna/visual_genome

Visual Genome



- Scene graphs

Language Grounding: Referring Expression Comprehension

Underlying Unimodal Tasks and Challenges

- Referring expression comprehension (+ generation)

Goal Situated language comprehension, including language disambiguation

Which person is meant?

guy in pink

Language Grounding: Referring Expression Comprehension

Why Relevant?

- Tests the joint comprehension of language and vision; fundamental to human communication

?? Dataset issues: language or image too simple, data bias, ...

RefCOCOg Val

a large green tractor in a parking lot

Language Grounding: Referring Expression Comprehension

Underlying Unimodal Tasks and Challenges

- Referring expression comprehension (+ generation)

Goal Situated language comprehension, including language disambiguation

Why Relevant?

- Tests the joint comprehension of language and vision; fundamental to human communication

?? Dataset issues: language or image too simple, data bias, ...

- Applications: Underlies basically all applications

Outline

Retrieval

Image–Text Matching

- Image–to–text retrieval: Given an image query, find the semantically closest text.
- Text–to–image retrieval: Given an text query, find the semantically closest image.
- Core approach: Compute the similarity between image and text

Retrieval

Image–Text Matching

- Image–to–text retrieval: Given an image query, find the semantically closest text.
- Text–to–image retrieval: Given an text query, find the semantically closest image.
- Core approach: Compute the similarity between image and text
- Popular dataset: Flickr30k [Young et al., 2014]

"A man sits in a chair while holding a large stuffed animal of a lion."
"A man is sitting on a chair holding a large stuffed animal."
"A man completes the finishing touches on a stuffed lion."
"A man holds a large stuffed lion toy."
"A man is smiling at a stuffed lion."

Outline

Relationship Recognition in Visuals

Several L&V tasks: Scene graph parsing; Visual semantic role labeling; Situation recognition

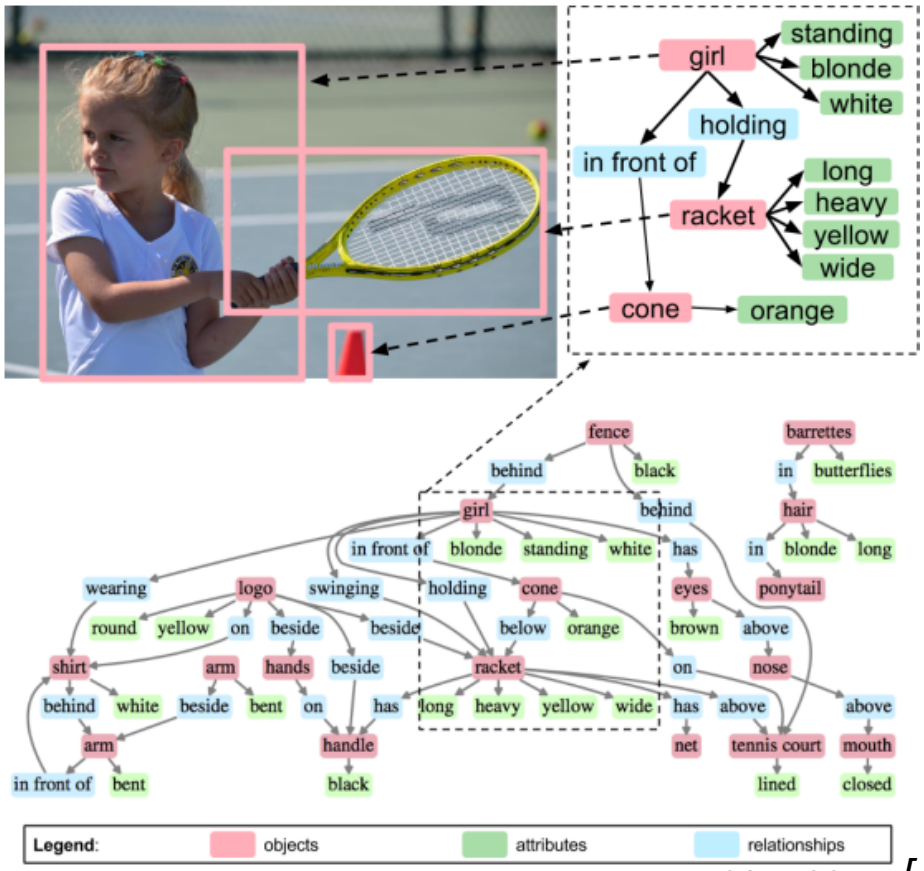
(a) Results for the query on a popular image search engine.

(b) Expected results for the query.

source: *[Johnson et al.,'15]*

Relationship Recognition in Visuals

Comprises several L&V tasks: *Scene graph parsing*; Visual semantic role labeling; Situation recognition



source: *[Johnson et al.,'15]*

Relationship Recognition in Visuals

Comprises several L&V tasks: Scene graph parsing; Visual semantic role labeling; *Situation recognition*

CLIPPING			
ROLE	VALUE	ROLE	VALUE
AGENT	MAN	AGENT	VET
SOURCE	SHEEP	SOURCE	DOG
TOOL	SHEARS	TOOL	CLIPPER
ITEM	WOOL	ITEM	CLAW
PLACE	FIELD	PLACE	ROOM

source: *[Yatskar et al.,'16;Pratt et al.,'20]*

Relationship Recognition in Visuals

Comprises several L&V tasks: Scene graph parsing; *Visual semantic role labeling*; Situation recognition

ARREST		PLACING	
r_1, r_2	Authorities	r_1	Agent
r_5	Suspect	r_5	Theme
r_3	Place	r_3	Place
		r_4	Goal

source: *[Silberer & Pinkal,'18]*

Relationship Recognition

Scene graph parsing; Visual semantic role labeling; Situation recognition

Underlying Unimodal Tasks and Challenges

- Text-based semantic role labeling and frame semantics $\Rightarrow$ semantic language understanding

Goal Deeper (semantic) image understanding

Why Relevant?

- Tests the semantic comprehension/recognition of the image content: the objecs, their properties and relations, the depicted situation

?? Large annotation effort for (semi-)supervised methods

- Applications: Cross-modal reasoning; Improvements on basic tasks such as image retrieval, VQA, etc.

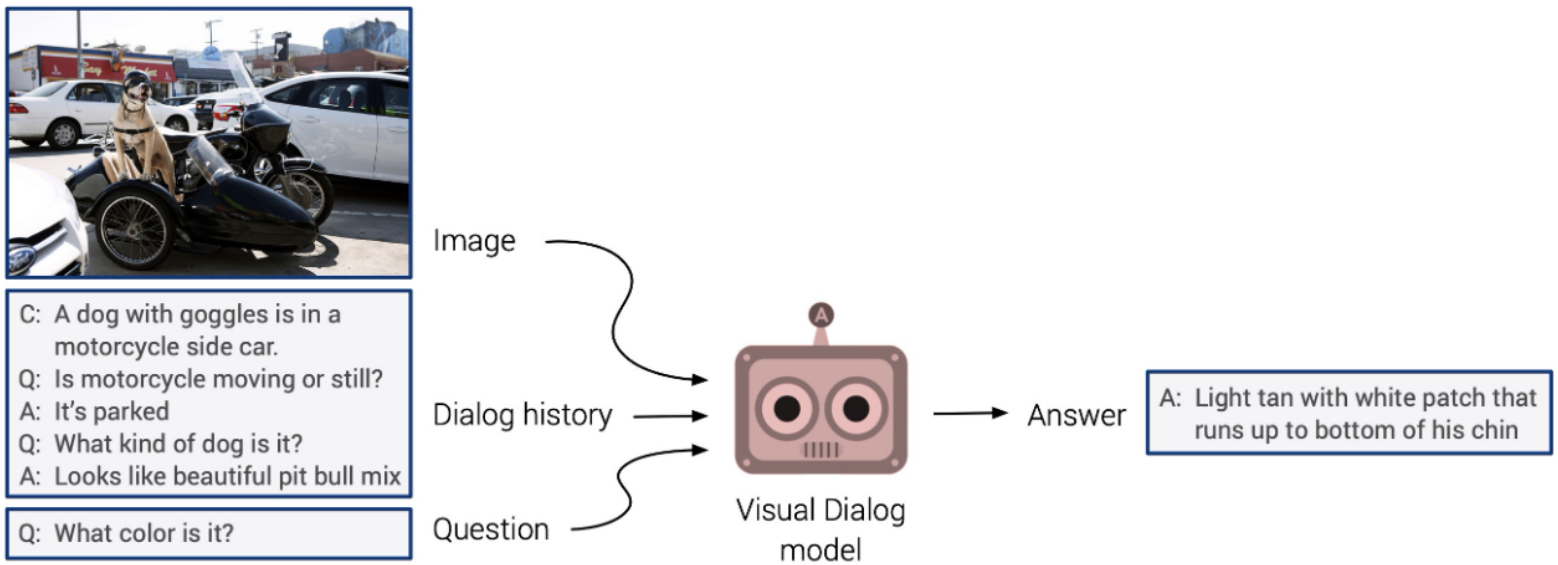
Visual Dialog

- Given an image, a dialog history, and a (follow-up) question about the image, answer the question

Visual Dialog

- Given an image, a dialog history, and a (follow-up) question about the image, answer the question

VisDial dataset

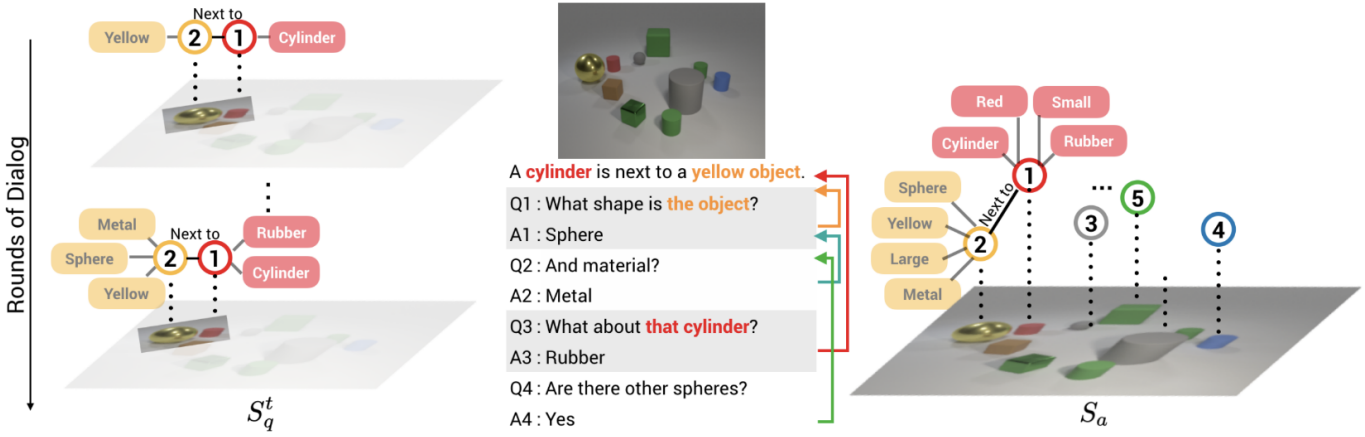


[Das et al., 2017]

Visual Dialog

- Given an image, a dialog history, and a (follow-up) question about the image, answer the question

CLEVR benchmark

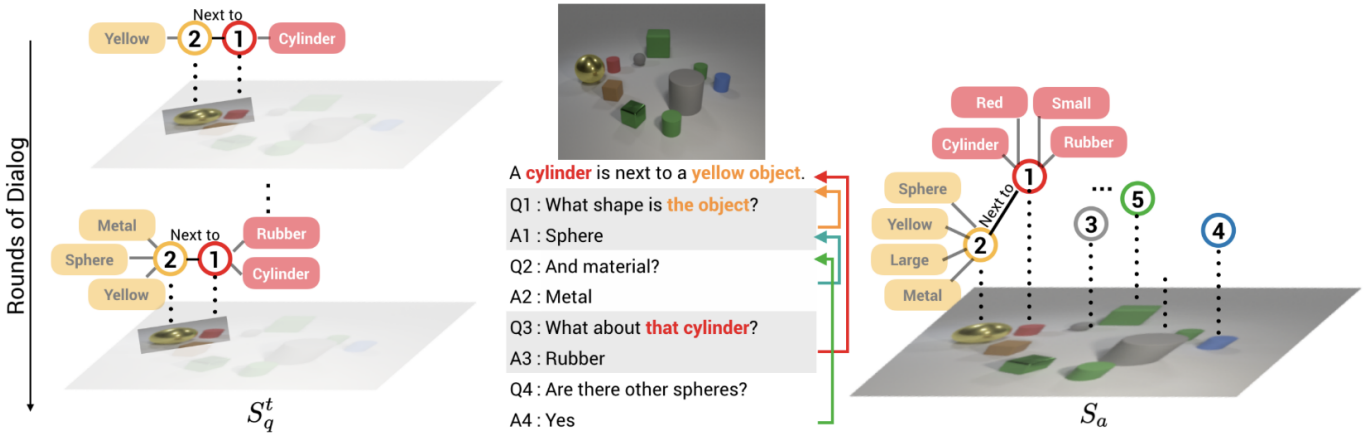


[Kottur et al., 2019]

Visual Dialog

- Given an image, a dialog history, and a (follow-up) question about the image, answer the question

CLEVR benchmark



[Kottur et al., 2019]

- Challenge: An AI agent is required to hold a 'meaningful' conversation with humans in natural language about visual content
- Underlying tasks: coreference resolution, reasoning
- Follow-up: [Wen et al., 2023] (long free-form answers)

Visual Entailment

https://github.com/necla-ml/SNLI-VE

SNLI-VE [Xie et al., 2018, Xie et al., 2019]

Given the evidence in the image:

Is "A couple are enjoying themselves at a bar." *true?*

Visual Entailment

https://github.com/necla-ml/SNLI-VE

SNLI-VE [Xie et al., 2018, Xie et al., 2019]

- Given an image as *premise* (P_{img}), and a natural language sentence as *hypothesis* (H_{text})
- Assign one of three labels based on the relationship conveyed by (P_{img}, H_{text}):
  - **entailment**: there is enough evidence in P_{img} to conclude that H_{text} is true
  - **contradiction**: there is enough evidence in P_{img} to conclude that H_{text} is false
  - **neutral** otherwise: the evidence in P_{img} is insufficient to draw a conclusion about H_{text}

Visual Entailment

https://github.com/necla-ml/SNLI-VE

SNLI-VE

[Xie et al., 2018, Xie et al., 2019]

A couple are enjoying themselves at a bar
-> entailment
A woman and a man meet for the first time on a date at the pub.
-> neutral
A group of people are riotously drunk in a bar
-> contradiction

Visual Entailment

https://nlp.stanford.edu/projects/snli/

Challenge in SNLI-VE

Fine-grained visual understanding—reason about the relationship between P_{img} and H_{text}

Underlying NLP Task: Natural Language Inference / Recognising Textual Entailment

- Given two short, ordered texts, determine the inference relation between them: entailment, contradiction, or neutral

Premise	Relation	Hypothesis
A black race car starts up in front of a crowd of people.	C	A man is driving down a lonely road.
An older and younger man smiling.	N	Two men are smiling and laughing at the cats playing on the floor.
A soccer game with multiple males playing.	E	Some men are playing a sport.

Visual Reasoning

GQA: A new dataset for real-world visual reasoning and compositional question answering [Hudson and Manning, 2019]

GQA: Examples

A1. Is the **tray** on top of the **table** black or light brown? light brown
A2. Are the **napkin** and the **cup** the same color? yes
A3. Is the small **table** both oval and wooden? yes
A4. Is there any **fruit** to the left of the **tray** the **cup** is on top of? yes
A5. Are there any **cups** to the left of the **tray** on top of the **table**? no
B1. What is the brown **animal** sitting inside of? **box**
B2. What is the large **container** made of? cardboard
B3. What **animal** is in the **box**? **bear**
B4. Is there a **bag** to the right of the green **door**? no
B5. Is there a **box** inside the plastic **bag**? no

Visual Reasoning

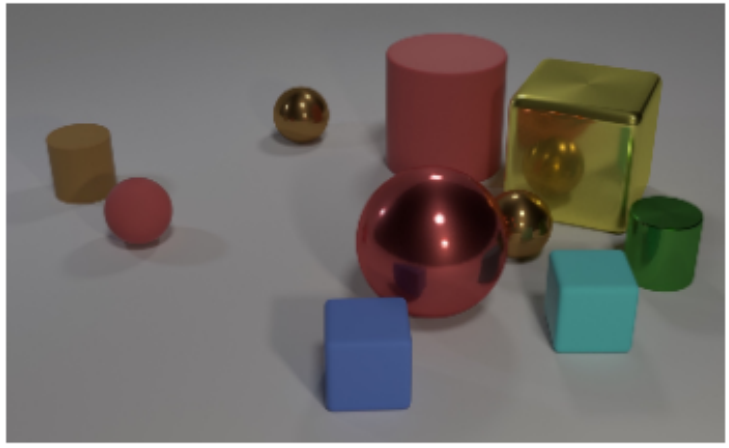
GQA: A new dataset for real-world visual reasoning and compositional question answering [Hudson and Manning, 2019]

GQA

- Goal: Address issues prevalent in previous VQA datasets:
  - real-world bias
  - linguistically and semantically simple questions (mere object recognition)
  - requires both linguistic and visual comprehension

Relational Reasoning

*C*LEVR: Compositional Language and Elementary Visual Reasoning



Q: Are there an equal number of large things and metal spheres?
Q: What size is the cylinder that is left of the brown metal thing that is left of the big sphere? **Q:** There is a sphere with the same size as the metal cube; is it made of the same material as the small red sphere?
Q: How many objects are either small cylinders or metal things?

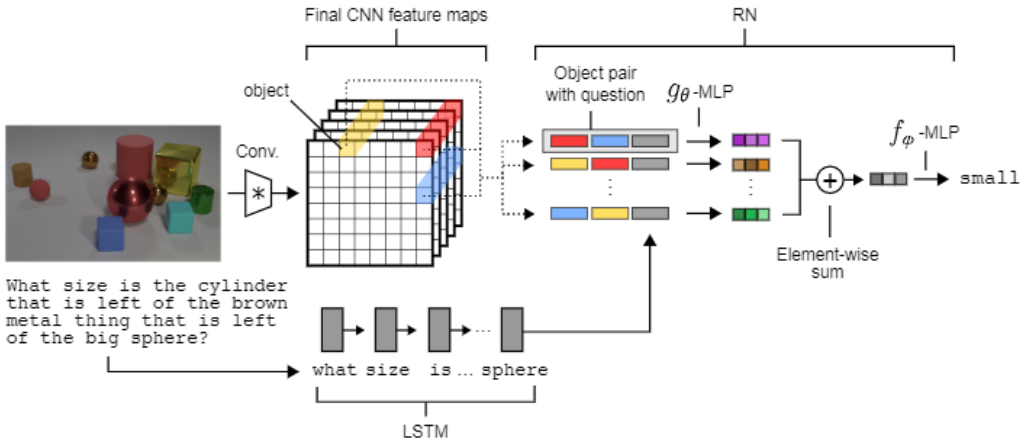
Figure 1. A sample image and questions from CLEVR. Questions test aspects of visual reasoning such as attribute identification, counting, comparison, multiple attention, and logical operations.

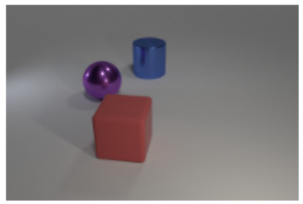
source: *[Johnson et al.,'16]*

Videos: CLEVRER [Yi et al., 2020]

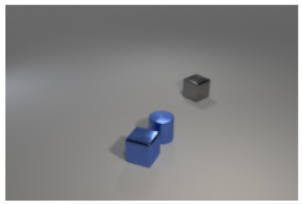
Relational Reasoning

A simple neural network module for relational reasoning

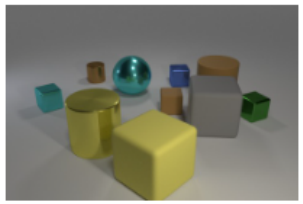




What number of objects are either big things that are left of the cylinder or cylinders?



Are there the same number of small blue objects that are to the right of the blue cube and blue metal cubes?



What number of other things are there of the same material as the green cube?

RN:	2	no	6
GT:	3	yes	5

source: *[Santoro et al.,'17]*

Visual Commonsense Reasoning

https://visualcommonsense.com [Zellers et al., 2019]

From Recognition to Cognition: Visual Commonsense Reasoning

- Task: Given an image, a list of regions, and a question, answer the question and provide a rationale
- Goal: Higher-order cognitive and commonsense understanding of the world

Why is [person4] pointing at [person1]?

a) He is telling [person3] that [person1] ordered the pancakes.
b) He just told a joke.
c) He is feeling accusatory towards [person1].
d) He is giving [person1] directions.

I chose a) because...

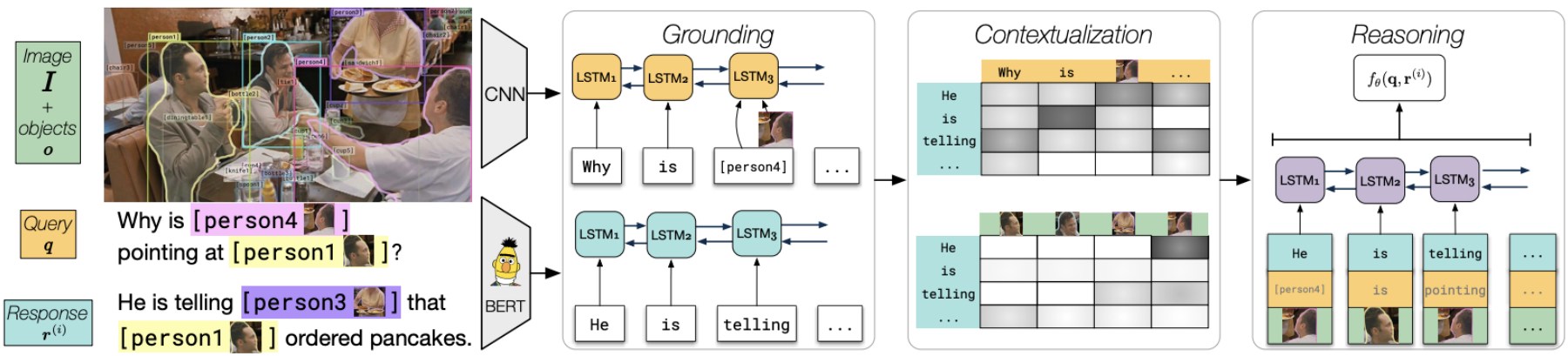
a) [person1] has the pancakes in front of him.
b) [person4] is taking everyone's order and asked for clarificat
c) [person3] is looking at the pancakes and both she and [person2] are smiling slightly.
d) [person3] is delivering food to the table, and she might n know whose order is whose.

Visual Commonsense Reasoning

https://visualcommonsense.com [Zellers et al., 2019]

From Recognition to Cognition: Visual Commonsense Reasoning

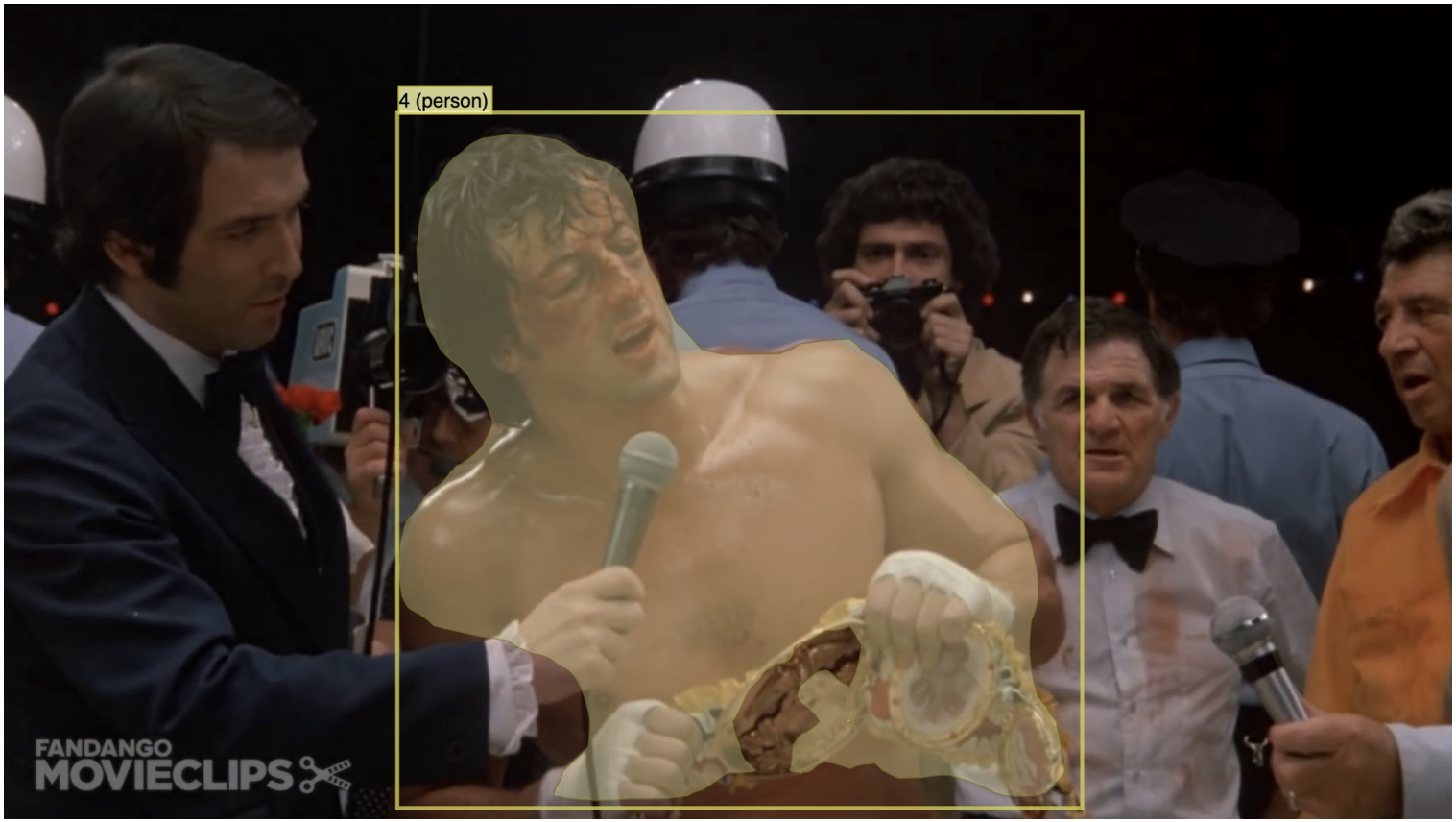
- Task: Given an image, a list of regions, and a question, answer the question and provide a rationale
- Goal: Higher-order cognitive and commonsense understanding of the world



Visual Commonsense Reasoning

https://visualcomet.xyz [Park et al., 2020]

What happened before / after; What is 4's intent?



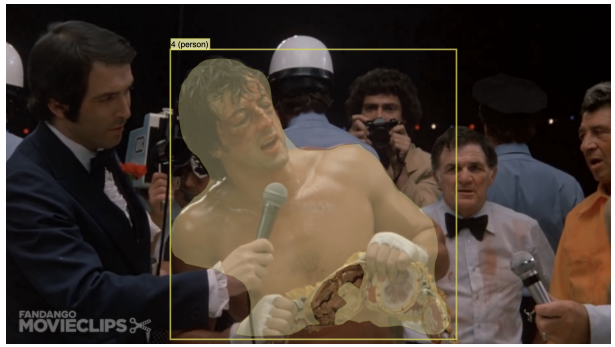
Place: in a boxing arena

Event: 4 is a boxer holding a title belt and talking into a microphone

Visual Commonsense Reasoning

https://visualcomet.xyz [Park et al., 2020]

What happened before / after; What is 4's intent?



Place: in a boxing arena

Event: 4 is a boxer holding a title belt and talking into a microphone

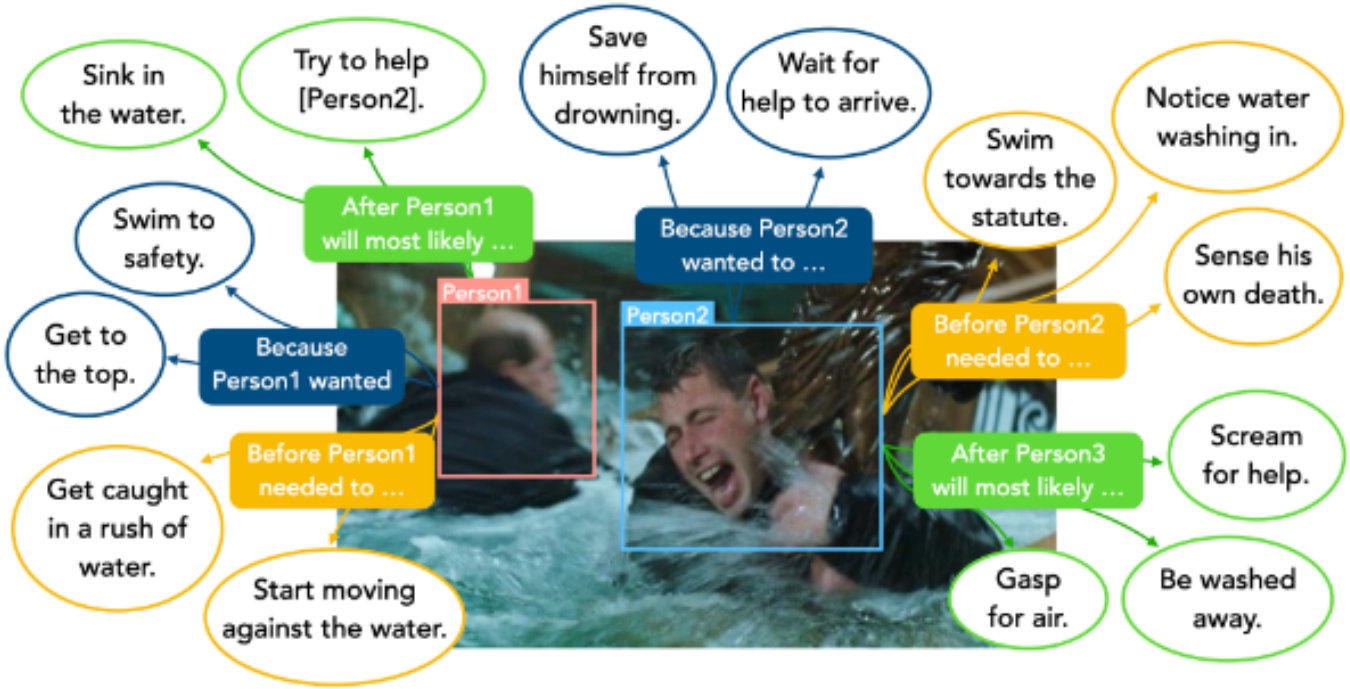
Before, PersonX needed to...	Because, PersonX wanted to...	After, PersonX will most likely...
-train to become the best boxer -put on his boxing gloves -size up his opponent -be punched in the face -get injured -win the match	-celebrate his victory -conduct a victory interview	-thank the audience -feel proud of himself -feel pain from being hit -lift up the belt -be cheered on by the crowd

Visual Commonsense Reasoning

https://visualcomet.xyz [Park et al., 2020]

VisualCOMET: Reasoning about the Dynamic Context of a Still Image

- Task: Given an image and context (textual description of the event at present / location of the image)
  Generate a set of commonsense inferences on
  - events before and after and
  - people's intents at present



Given a person in the image, VisualCOMET provides a *graph* of

Visual Commonsense Reasoning

Underlying Unimodal Tasks and Challenges

- Object recognition and grounding
- (Relational) Reasoning in text

Goal Situated compositional language understanding; Reason about the visual scene / story

Why Relevant?

- Tests joint cognitive understanding of language and visuals using background (commonsense) knowledge; fundamental to human cognition
- Applications: =)

Current Trends in CV: Image Generation, Editing, Impainting

Text-to-image generation

"a hedgehog using a calculator"

"a corgi wearing a red bowtie and a purple party hat"

"robots meditating in a vipassana retreat"

"a fall landscape with a small cottage next to a lake"

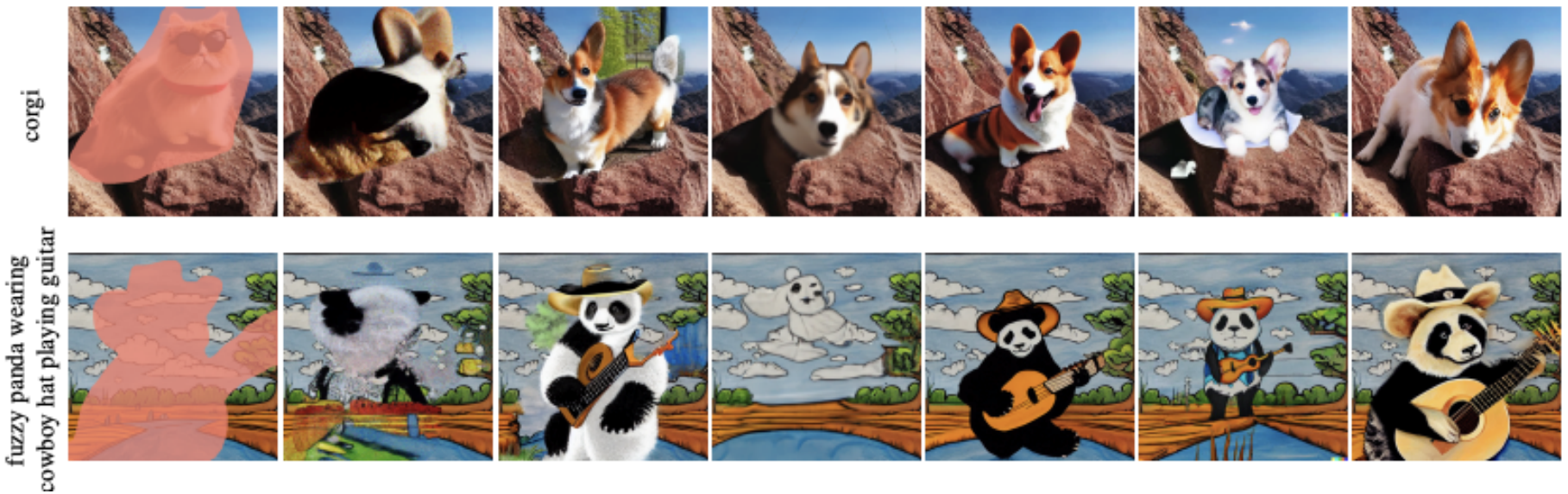
[Zhang et al., 2023a]

Stable Diffusion demo: `https://huggingface.co/spaces/stabilityai/stable-diffusion`

Current Trends in CV: Image Generation, Editing, Impainting

Image impainting with mask + text

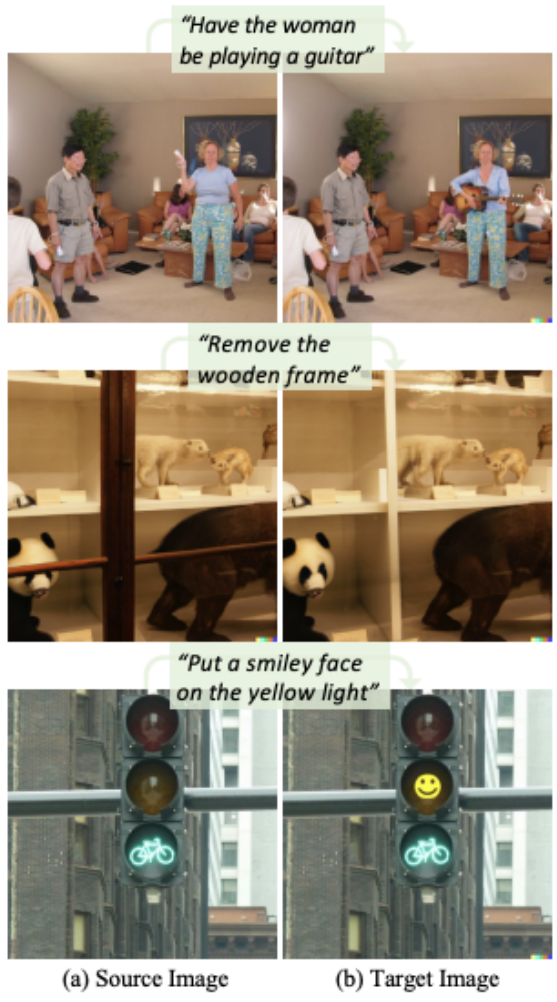
GLIDE model



[Xie et al., 2023]

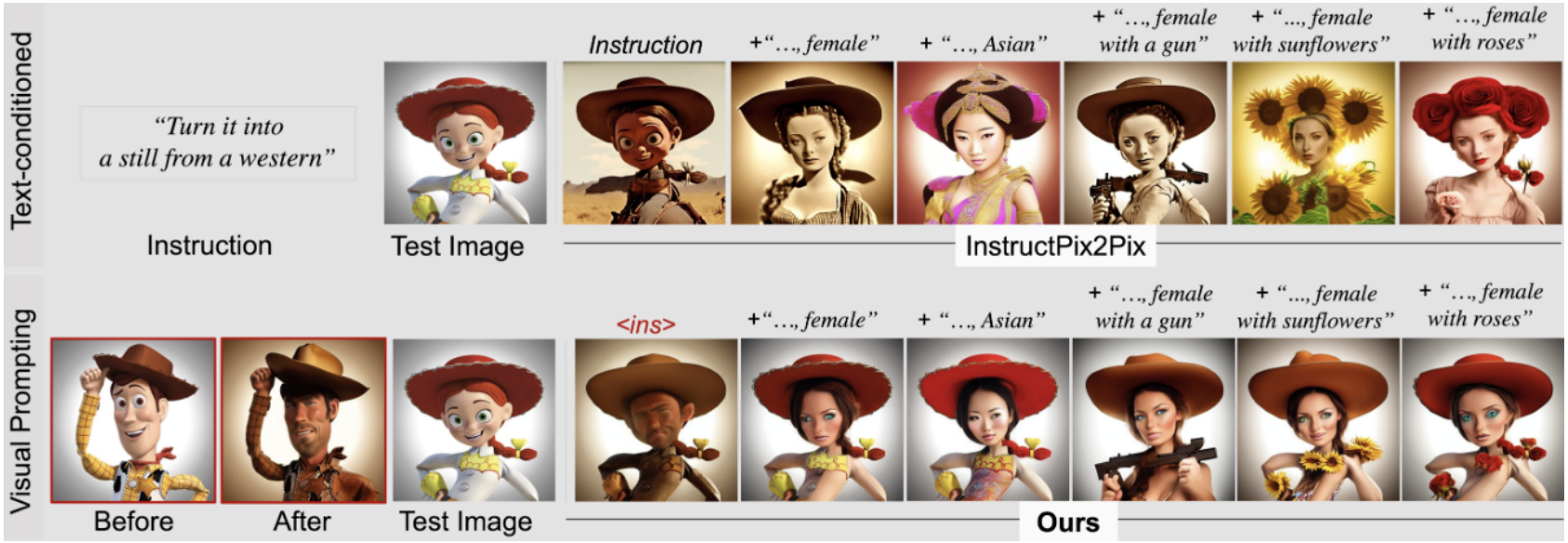
Current Trends in CV: Image Generation, Editing, Impainting

Instruction-guided image editing



Current Trends in CV: Image Generation, Editing, Impainting

Image editing with text + visual prompt



[Nguyen et al., 2023]

Current Trends in L+V: Explanations

Reasoning with Explanations / Rationales

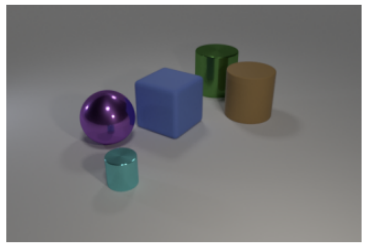
VQA-X
Question: Does this scene look like it could be from the early 1950s?

Answer | Explanation:
Yes | The photo is in black and white and the cars are all classic designs from the 1950s

e-SNLI-VE
Hypothesis: A woman is holding a child.

Answer | Explanation:
Entailment | If a woman holds a child she is holding a child.

CLEVR-X
Question: There is a purple metallic ball; what number of cyan objects are right of it?



Answer | Explanation:
1 | There is a cyan cylinder which is on the right side of the purple metallic ball.

Fig. 1: Comparing examples from the VQA-X (left), e-SNLI-VE (middle), and CLEVR-X (right) datasets. The explanation in VQA-X requires prior knowledge (about cars from the 1950s), e-SNLI-VE argues with a tautology, and our CLEVR-X only uses abstract visual reasoning.

References, Tools, Code

- https://github.com/salesforce/LAVIS
- https://www.visuallanguagelab.com/mast
- https://mmf.sh/docs/notes/projects/
  https://github.com/facebookresearch/mmf
- huggingface.co/spaces/timbrooks/instruct-pix2pix
- What is multimodality? [Parcalabescu et al., 2021]

References I

Bisk, Y., Holtzman, A., Thomason, J., Andreas, J., Bengio, Y., Chai, J., Lapata, M., Lazaridou, A., May, J., Nisnevich, A., et al. (2020).
Experience grounds language.
In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8718–8735.

Das, A., Kottur, S., Gupta, K., Singh, A., Yadav, D., Moura, J. M., Parikh, D., and Batra, D. (2017).
Visual dialog.
In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 326–335.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. (2021).
An image is worth 16x16 words: Transformers for image recognition at scale.
In *International Conference on Learning Representations*.

Hudson, D. A. and Manning, C. D. (2019).
GQA: A new dataset for real-world visual reasoning and compositional question answering.
In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709.

Kottur, S., Moura, J. M. F., Parikh, D., Batra, D., and Rohrbach, M. (2019).
CLEVR-dialog: A diagnostic dataset for multi-round reasoning in visual dialog.
In Burstein, J., Doran, C., and Solorio, T., editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 582–595, Minneapolis, Minnesota. Association for Computational Linguistics.

References II

Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.-J., Shamma, D. A., Bernstein, M. S., and Fei-Fei, L. (2017).
Visual genome: Connecting language and vision using crowdsourced dense image annotations.
Int. J. Comput. Vision, 123(1):32–73.

Liang, P. P., Zadeh, A., and Morency, L.-P. (2022).
Foundations and trends in multimodal machine learning: Principles, challenges, and open questions.
arXiv e-prints, pages arXiv–2209.

Liu, L., Ouyang, W., Wang, X., Fieguth, P., Chen, J., Liu, X., and Pietikäinen, M. (2020).
Deep learning for generic object detection: A survey.
International journal of computer vision, 128:261–318.

Mogadala, A., Kalimuthu, M., and Klakow, D. (2021).
Trends in integration of vision and language research: A survey of tasks, datasets, and methods.
J. Artif. Int. Res., 71:1183–1317.

Nakayama, H., Tamura, A., and Ninomiya, T. (2020).
A visually-grounded parallel corpus with phrase-to-region linking.
In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4204–4210, Marseille, France. European Language Resources Association.

Nguyen, T., Li, Y., Ojha, U., and Lee, Y. J. (2023).
Visual instruction inversion: Image editing via visual prompting.
In *Thirty-seventh Conference on Neural Information Processing Systems*.

References III

Parcalabescu, L., Trost, N., and Frank, A. (2021).

What is multimodality?

In Donatelli, L., Krishnaswamy, N., Lai, K., and Pustejovsky, J., editors, *Proceedings of the 1st Workshop on Multimodal Semantic Representations (MMSR)*, pages 1–10, Groningen, Netherlands (Online). Association for Computational Linguistics.

Park, J. S., Bhagavatula, C., Mottaghi, R., Farhadi, A., and Choi, Y. (2020).

Visualcomet: Reasoning about the dynamic context of a still image.

In *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V*, page 508–524, Berlin, Heidelberg. Springer-Verlag.

Plummer, B. A., Wang, L., Cervantes, C. M., Caicedo, J. C., Hockenmaier, J., and Lazebnik, S. (2015).

Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models.

In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 2641–2649.

Wang, P., Wu, Q., Cao, J., Shen, C., Gao, L., and van den Hengel, A. (2018).

Neighbourhood watch: Referring expression comprehension via language-guided graph attention networks.

2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 1960–1968.

Wen, B., Yang, Z., Wang, J., Gan, Z., Howe, B., and Wang, L. (2023).

Infovisdial: An informative visual dialogue dataset by bridging large multimodal and language models.

arXiv e-prints, pages arXiv–2312.

Xie, N., Lai, F., Doran, D., and Kadav, A. (2018).

Visual entailment task for visually-grounded language learning.

arXiv preprint arXiv:1811.10582.

References IV

Xie, N., Lai, F., Doran, D., and Kadav, A. (2019).

Visual entailment: A novel task for fine-grained image understanding.
arXiv preprint arXiv:1901.06706.

Xie, S., Zhang, Z., Lin, Z., Hinz, T., and Zhang, K. (2023).

Smartbrush: Text and shape guided object inpainting with diffusion model.
In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22428–22437, Los Alamitos, CA, USA. IEEE Computer Society.

Yi, K., Gan, C., Li, Y., Kohli, P., Wu, J., Torralba, A., and Tenenbaum, J. B. (2020).

Clevrer: Collision events for video representation and reasoning.
In *International Conference on Learning Representations*.

Young, P., Lai, A., Hodosh, M., and Hockenmaier, J. (2014).

From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions.
TACL, 2:67–78.

Zellers, R., Bisk, Y., Farhadi, A., and Choi, Y. (2019).

From recognition to cognition: Visual commonsense reasoning.
In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6720–6731.

Zhang, C., Zhang, C., Zhang, M., and Kweon, I. S. (2023a).

Text-to-image diffusion models in generative ai: A survey.

Zhang, K., Mo, L., Chen, W., Sun, H., and Su, Y. (2023b).

Magicbrush: A manually annotated dataset for instruction-guided image editing.
In *Advances in Neural Information Processing Systems*.