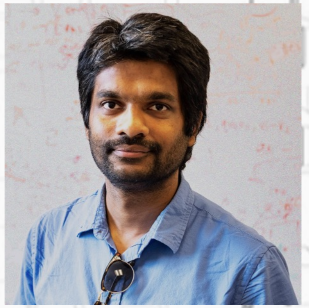
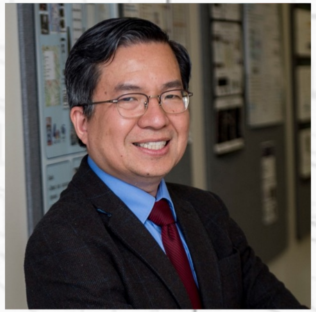
Knowledge-Driven
Vision-Language Pretraining


Manling Li
UIUC


Xudong Lin
Columbia


Jie Lei
Meta AI


Mohit Bansal
UNC


Shih-Fu Chang
Columbia


Heng Ji
UIUC

UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN
COLUMBIA UNIVERSITY
UNC
Meta AI

Why Vision-Language?

- Multimedia downstream tasks

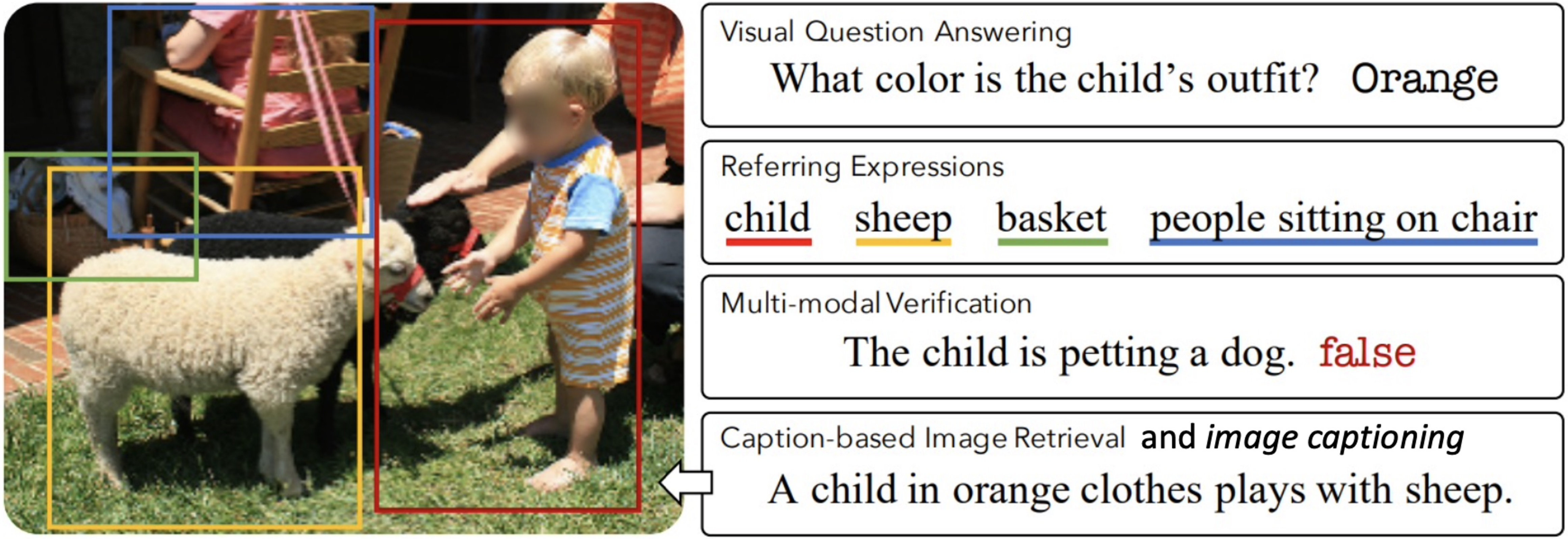
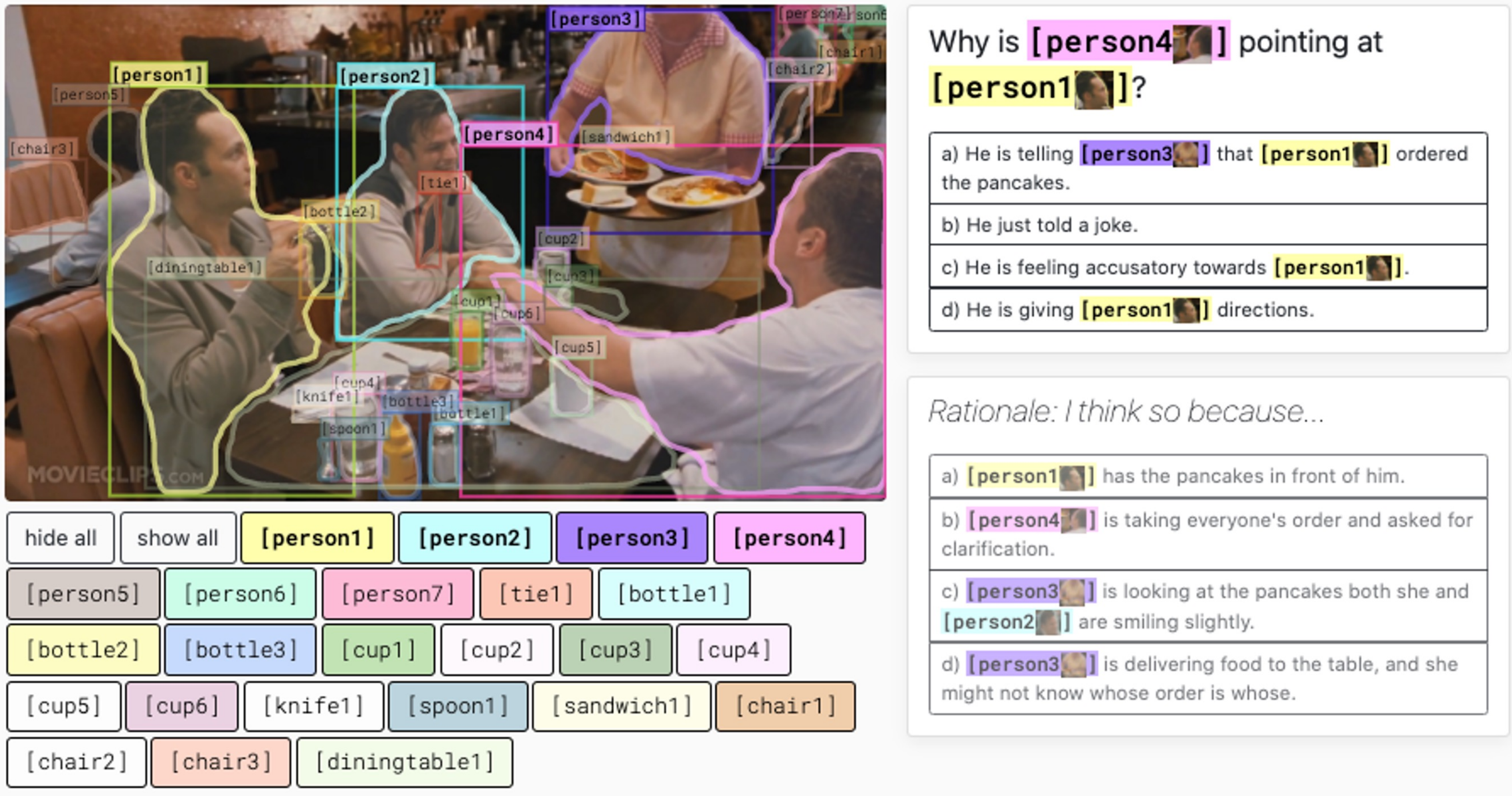


Image from: https://arxiv.org/pdf/1912.02315.pdf

Why Vision-Language?

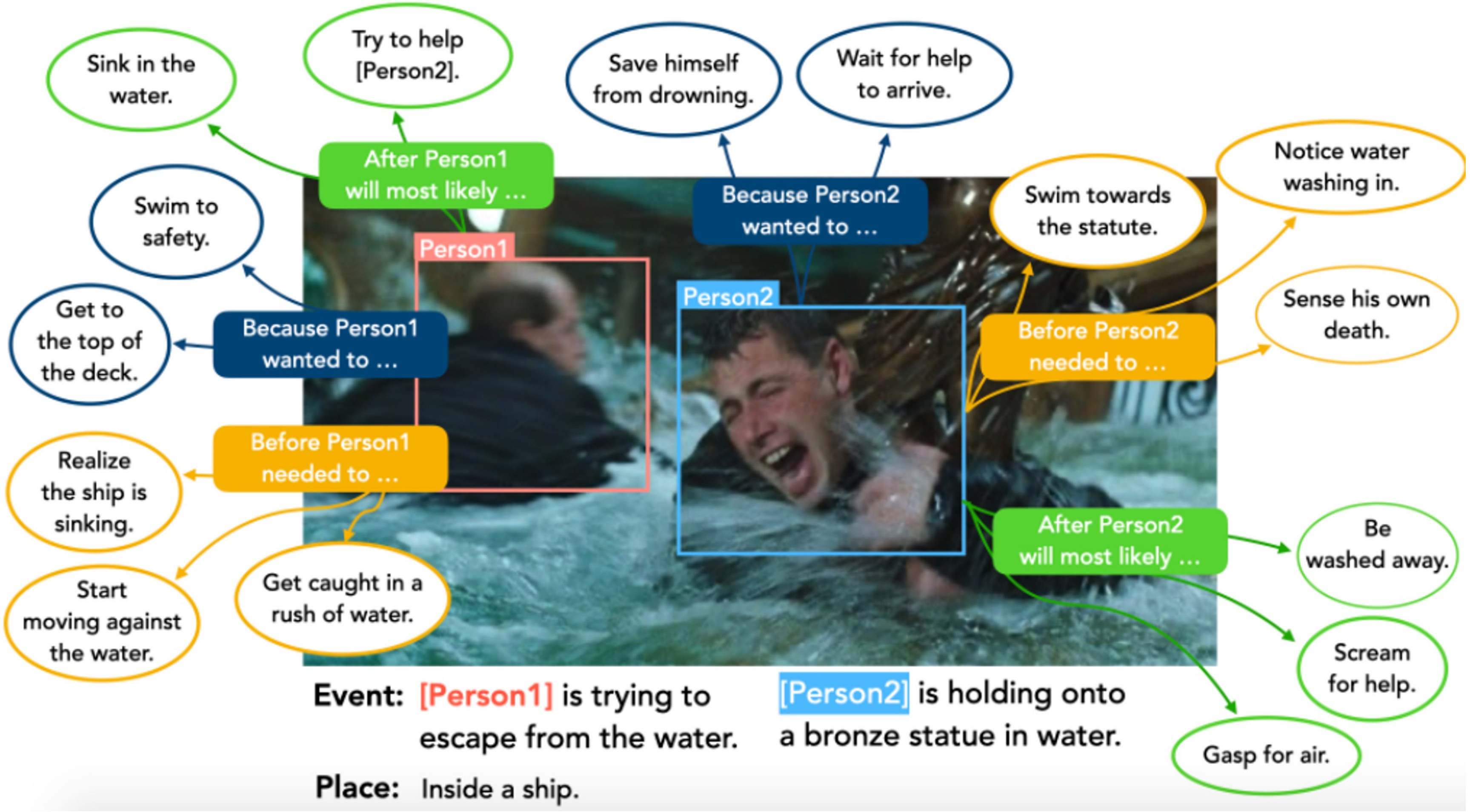
- VCR: Visual Commonsense Reasoning



Zellers, Rowan, et al. "From recognition to cognition: Visual commonsense reasoning." *CVPR*. 2019.

Why Vision-Language?

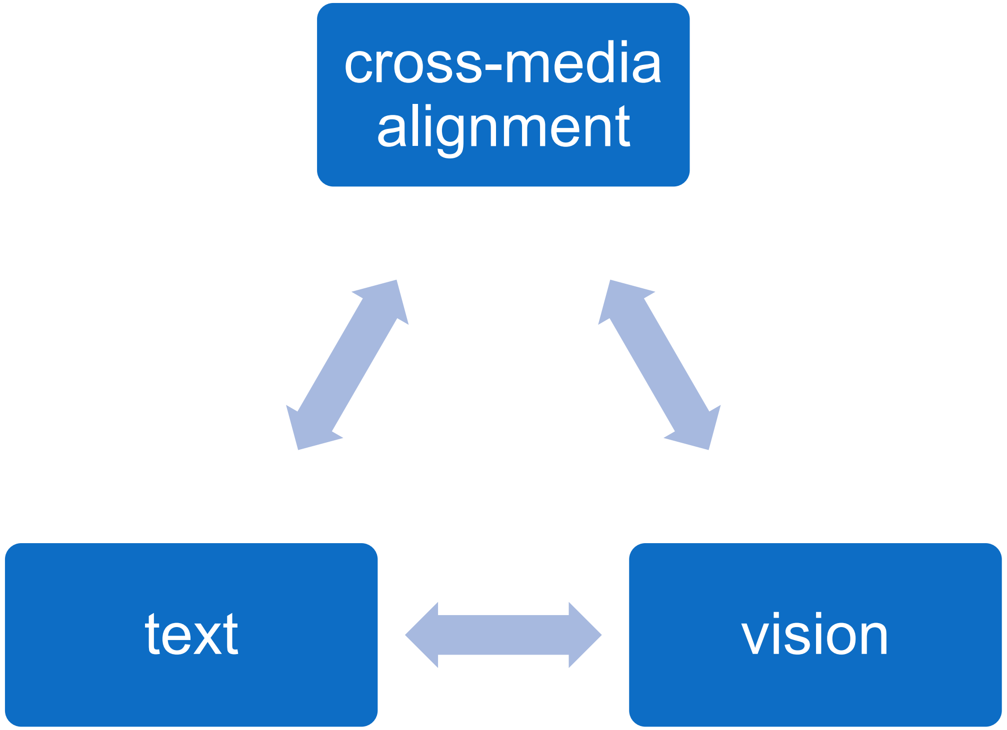
- VisualCOMET: Visual Commonsense Reasoning in Time



Park, Jae Sung, et al. "Visualcomet: Reasoning about the dynamic context of a still image." ECCV. Springer, Cham, 2020.

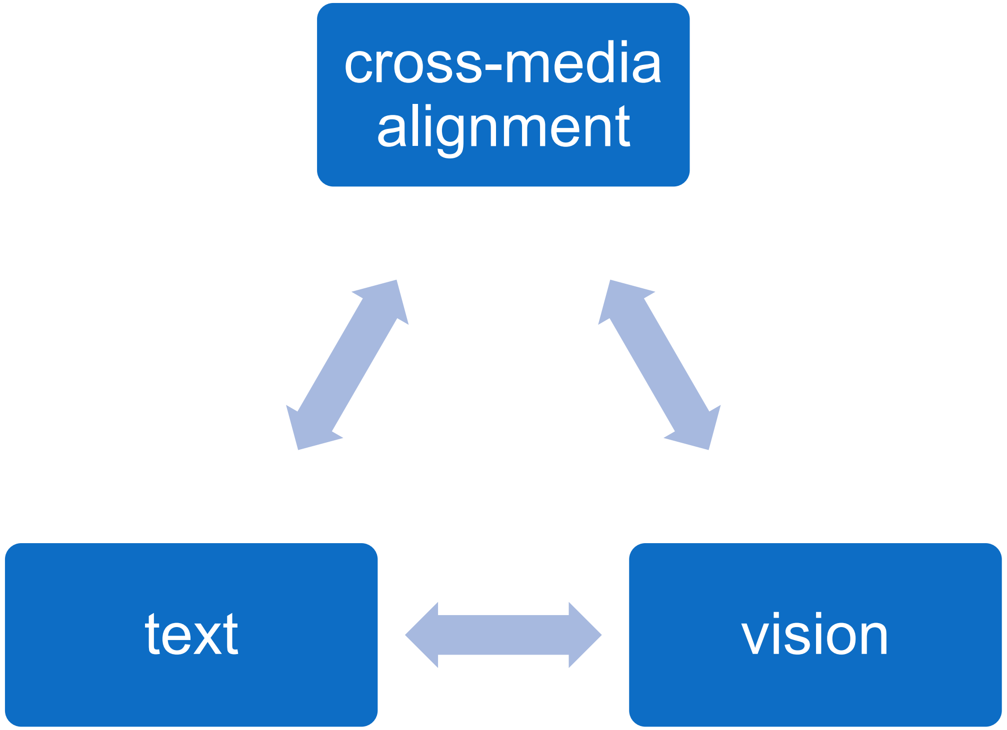
Architecture of Vision-Language Pretraining

- How? Vision-language pre-training by vision-language pairs



Architecture of Vision-Language Pretraining

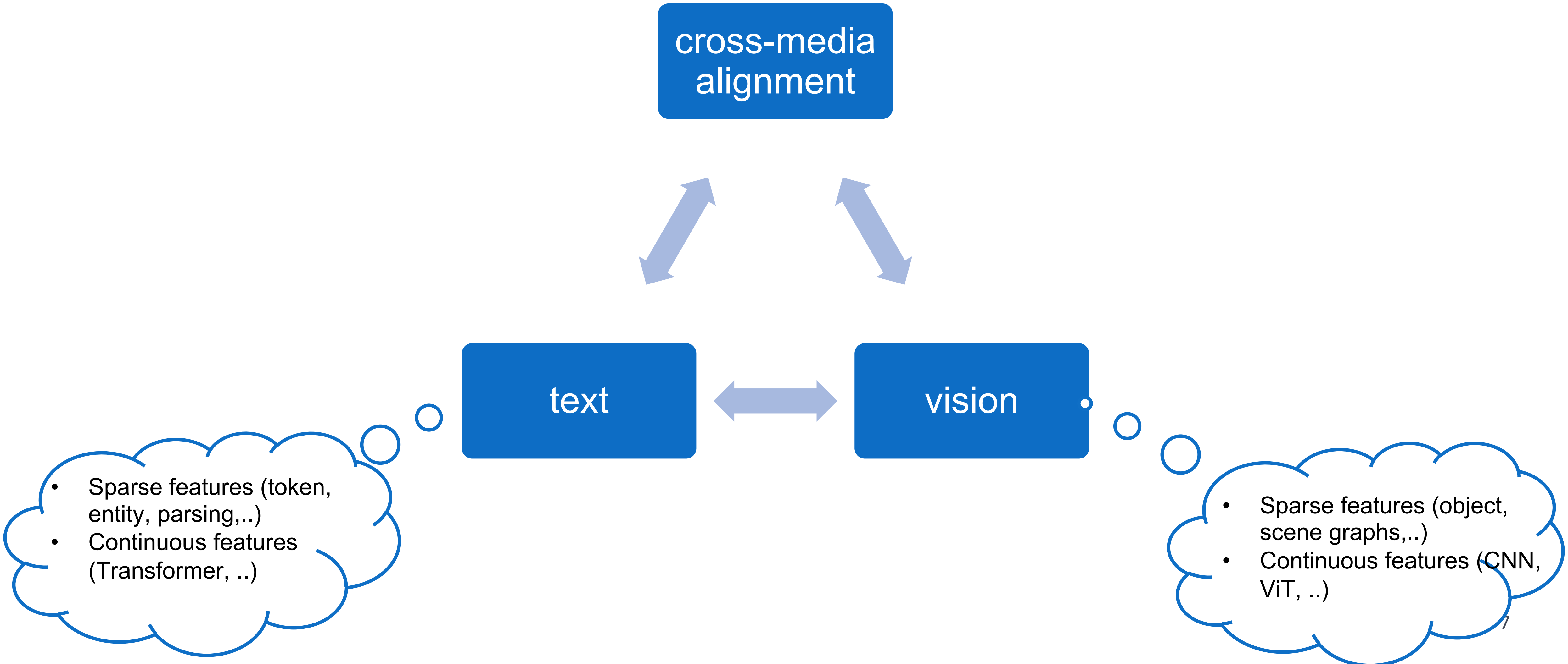
- How? Image-text pre-training by image-text pairs



- Sparse features (token, entity, parsing,..)
- Continuous features (Transformer, ..)

Architecture of Vision-Language Pretraining

- How? Image-text pre-training by image-text pairs



Typical Loss: Self-supervised in Vision

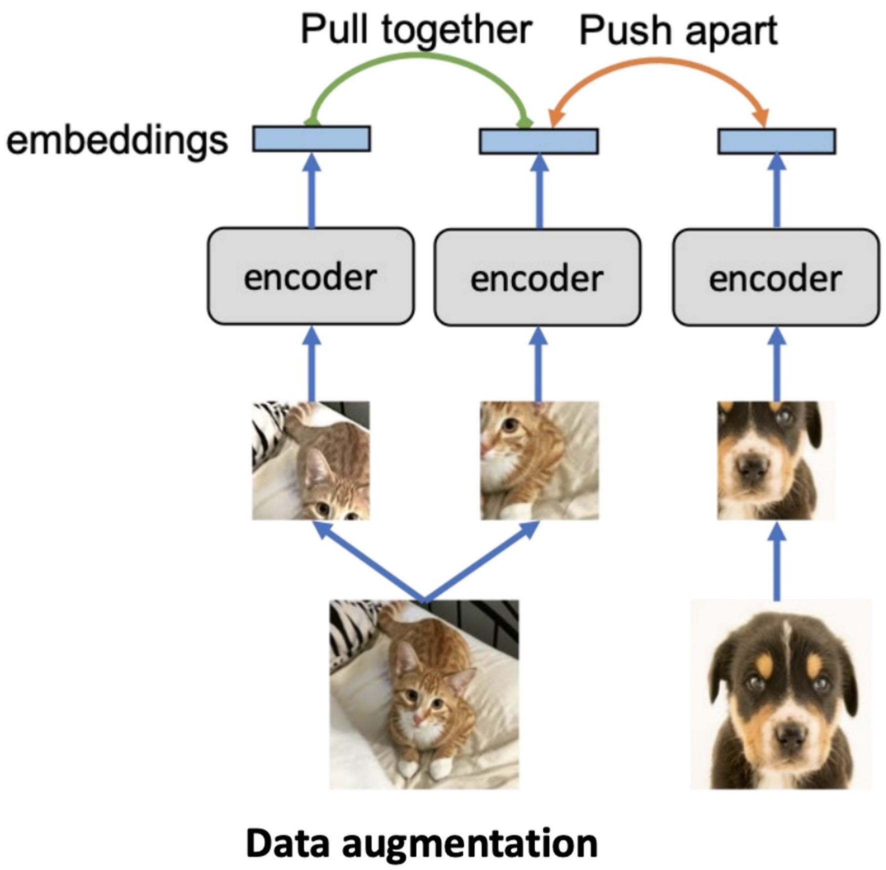
Contrastive Learning

- large batch size required
- data hungry
- hard example sensitive

Generative (Masked Prediction)

- Masked object prediction
- Masked feature regression

Self-supervision in Vision: Contrastive Learning



SimCLR (Chen et al., 2020), MoCo (He et al., 2020), DINO (Caron et al., 2021)

Contrastive Learning

- large batch size required
- data hungry
- hard example sensitive

Images from original authors

Typical Loss: Self-supervised in Vision

Contrastive Learning

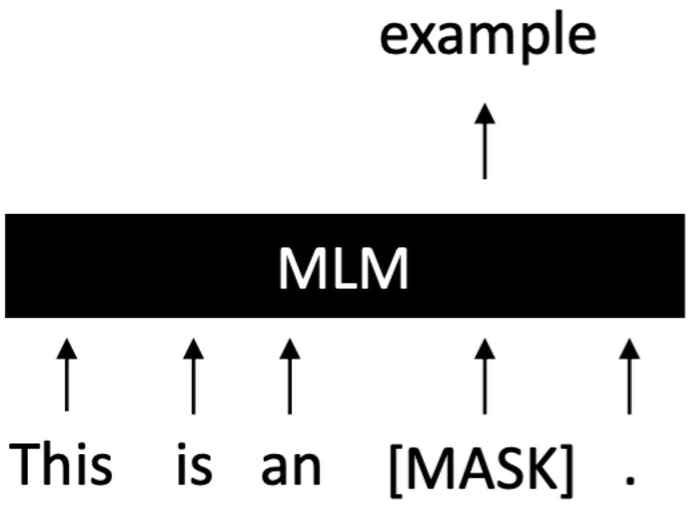
- large batch size required
- data hungry
- hard example sensitive

Generative (Masked Prediction)

- Masked feature regression
- Masked object prediction

Self-supervision in Vision: Masked Language Model

- Split an image into patches



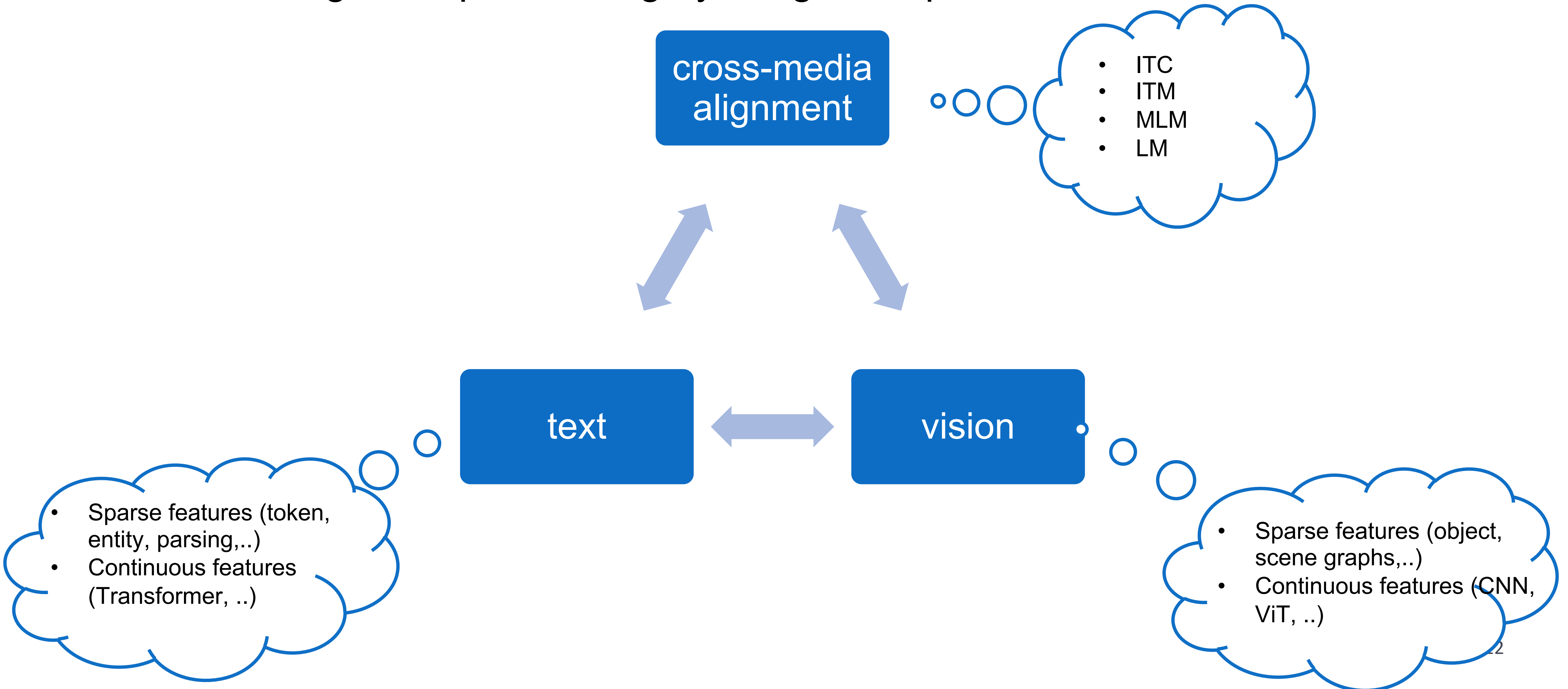
Masked **Image** Modeling

BERT Masked
Language Modeling

Slides from original authors

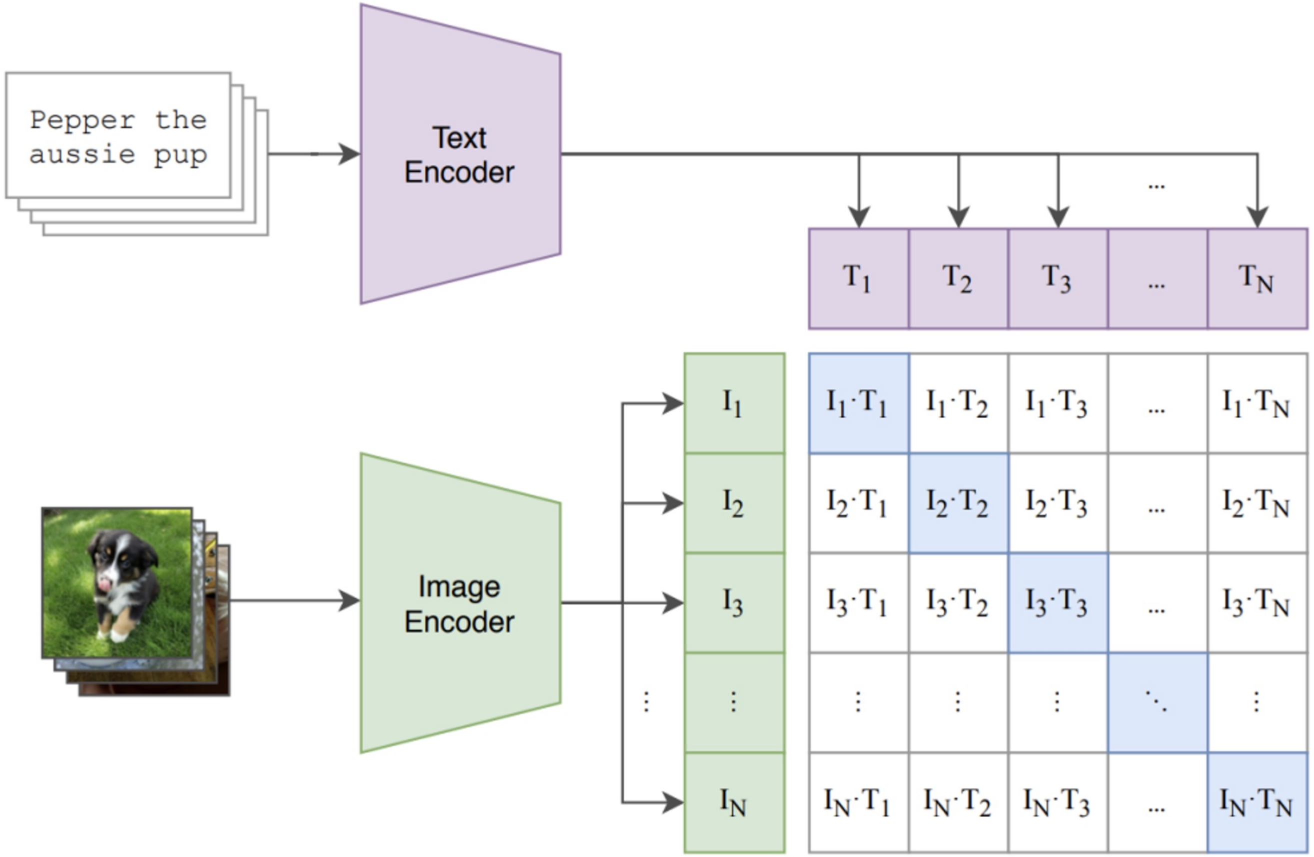
The Goal of Vision-Language Pretraining

- How? Image-text pre-training by image-text pairs



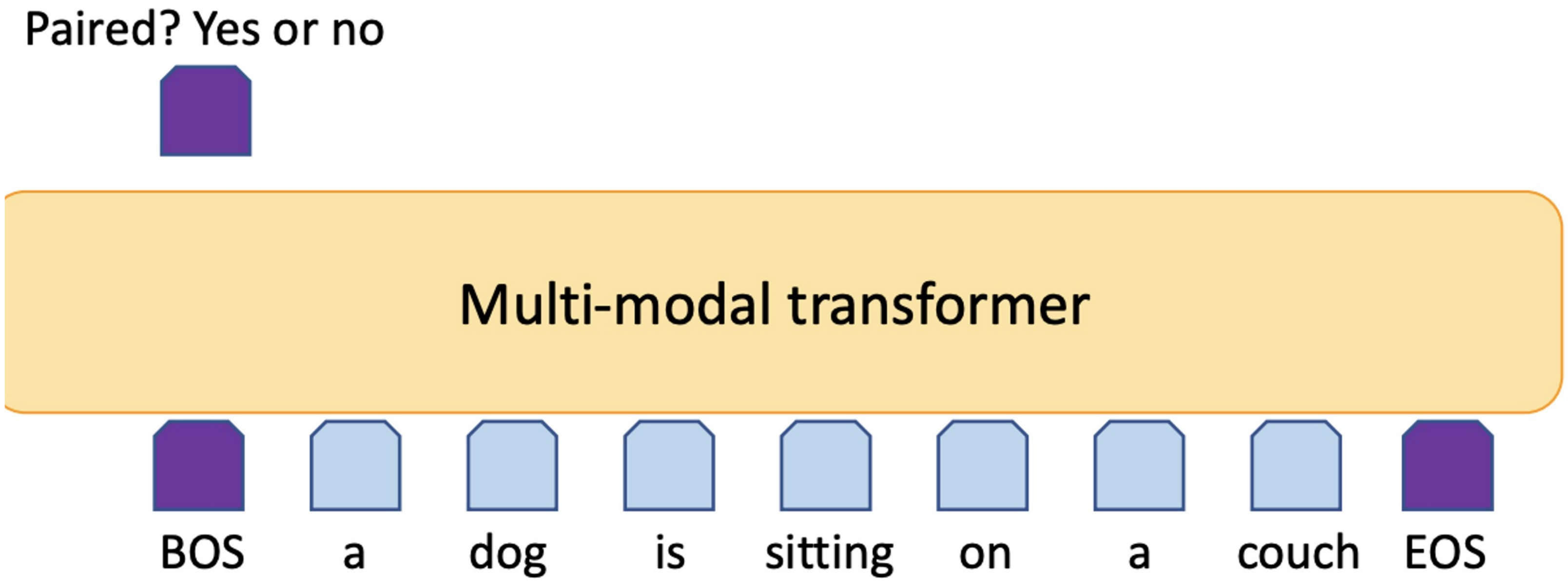
Typical Loss: Image-text contrastive (ITC)

- Image-text contrastive (ITC) loss



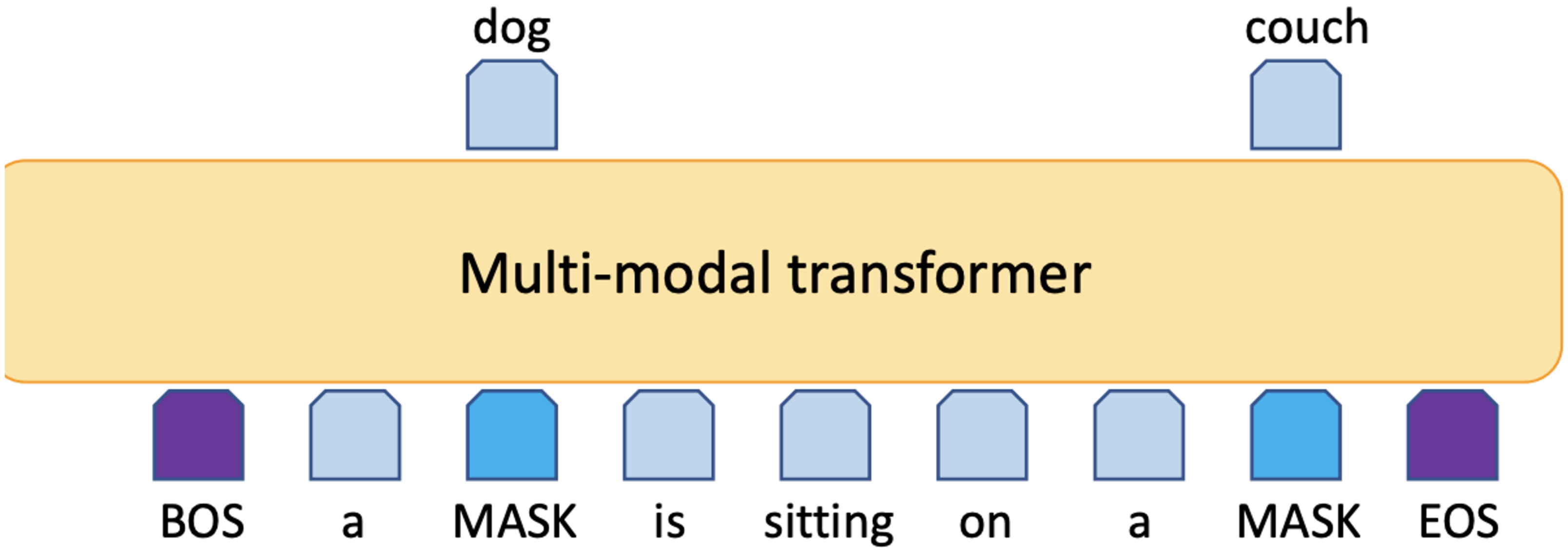
A Simple Framework for Contrastive Learning of Visual Representations, 2020
Momentum Contrast for Unsupervised Visual Representation Learning, 2019
Align before Fuse: Vision and Language Representation Learning with Momentum Distillation, 2021
UFO: A UniFied TransfOrmer for Vision-Language Representation Learning, 2021

Typical Loss: Image-text matching (ITM) loss



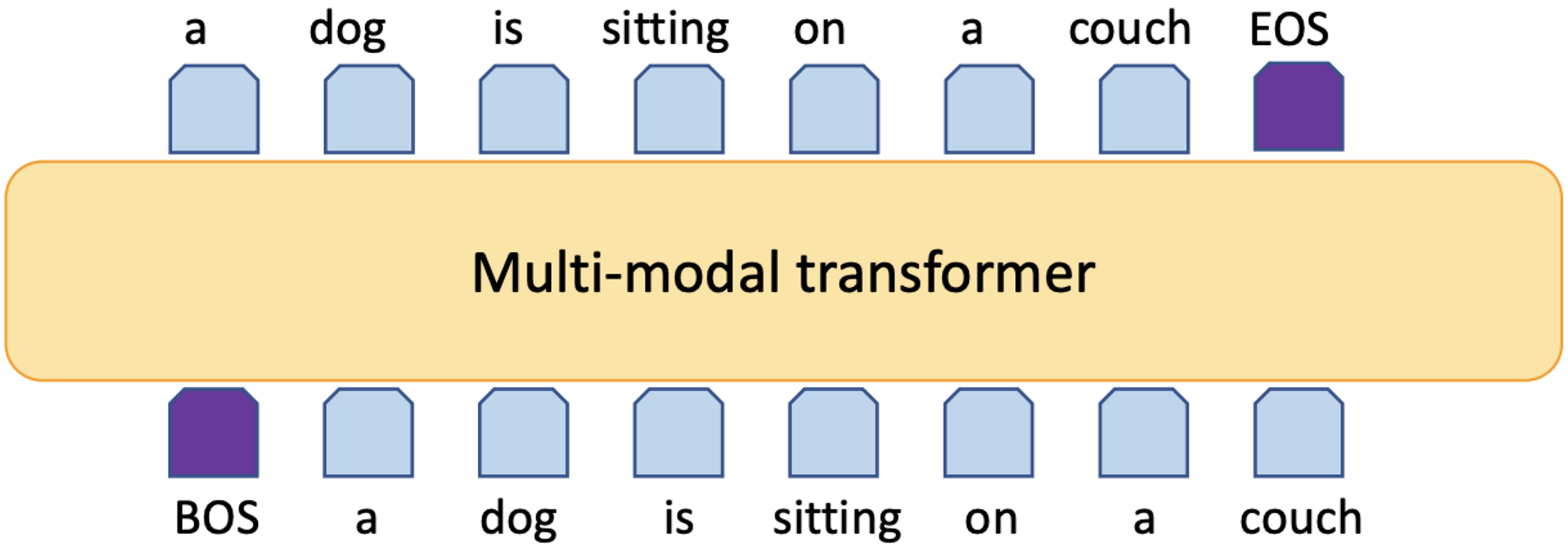
https://datarelease.blob.core.windows.net/tutorial/VLP-Tutorial_2022/image_text_part1.pdf

Typical Loss: Masked language modeling (MLM) loss



https://datarelease.blob.core.windows.net/tutorial/VLP-Tutorial_2022/image_text_part1.pdf

Typical Loss: Language modeling (LM) loss



https://datarelease.blob.core.windows.net/tutorial/VLP-Tutorial_2022/image_text_part1.pdf

Why Knowledge?

Compared to raw data, knowledge is **important and useful information.**

Text	Vision
Entity	Object
Relation	Scene Graph
Event	Activity/Situation

Factual Knowledge

Explicit Knowledge

Implicit Knowledge

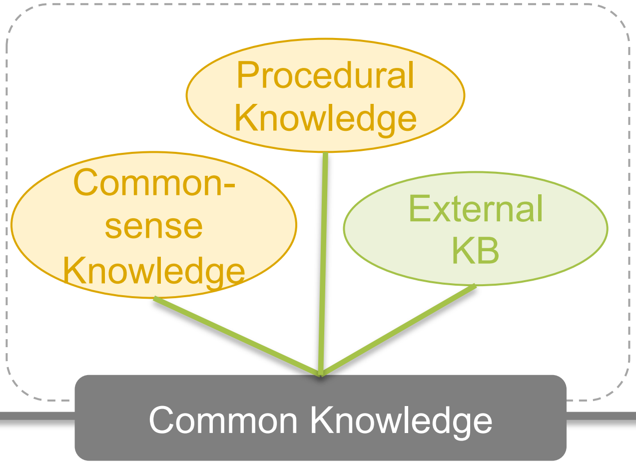
Why Knowledge?

Compared to raw data, knowledge is **important and useful information.**

Explicit Knowledge

Text	Vision
Entity	Object
Relation	Scene Graph
Event	Activity/Situation

Factual Knowledge



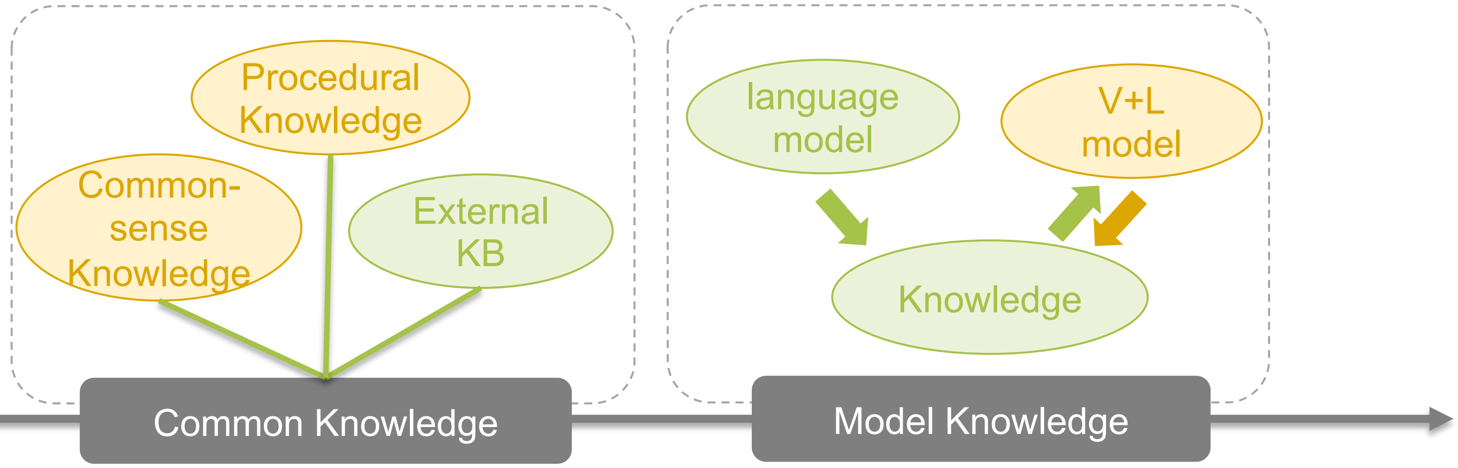
Implicit Knowledge

Why Knowledge?

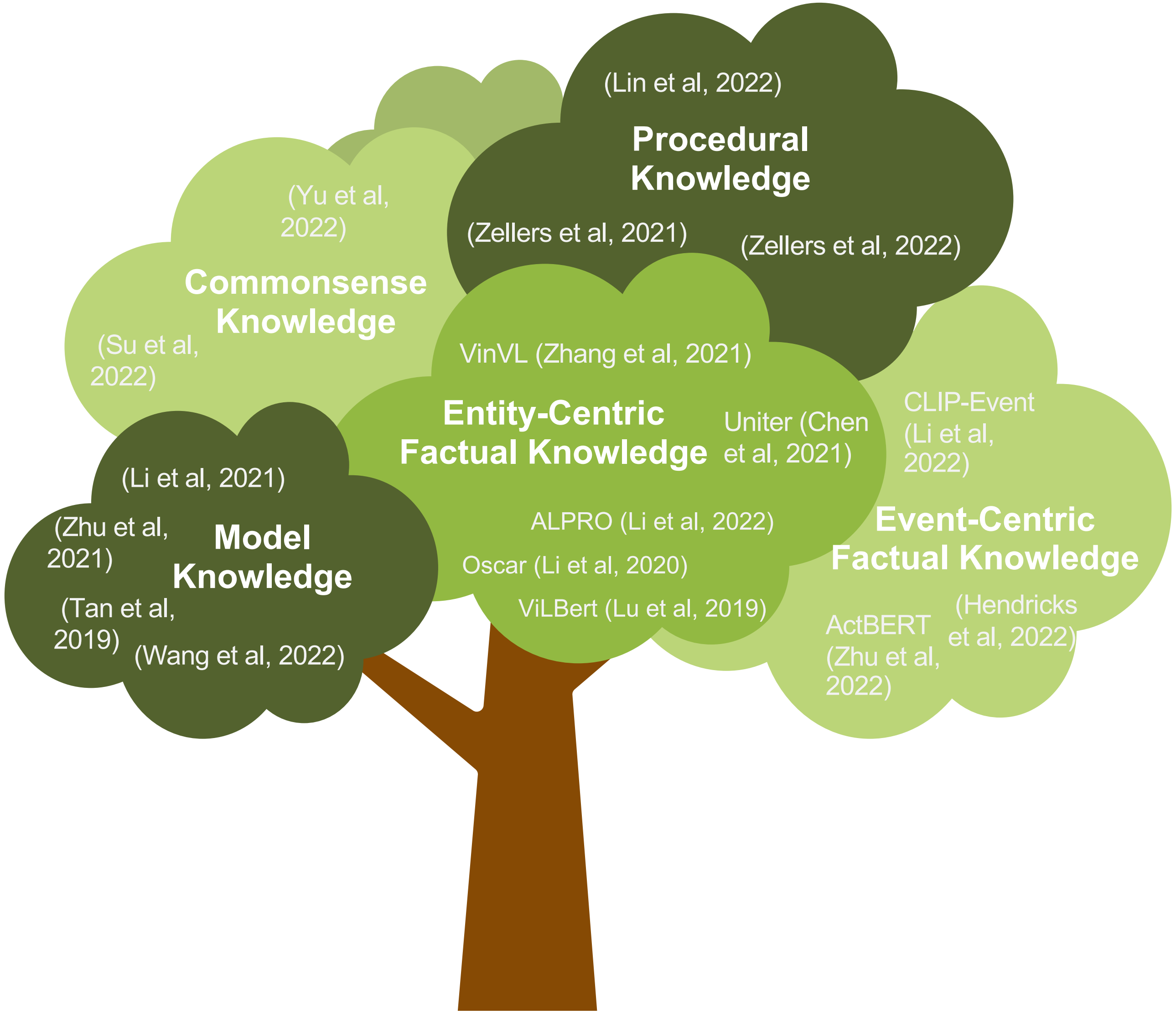
Compared to raw data, knowledge is **important and useful information.**

Text	Vision
Entity	Object
Relation	Scene Graph
Event	Activity/Situation

Factual Knowledge

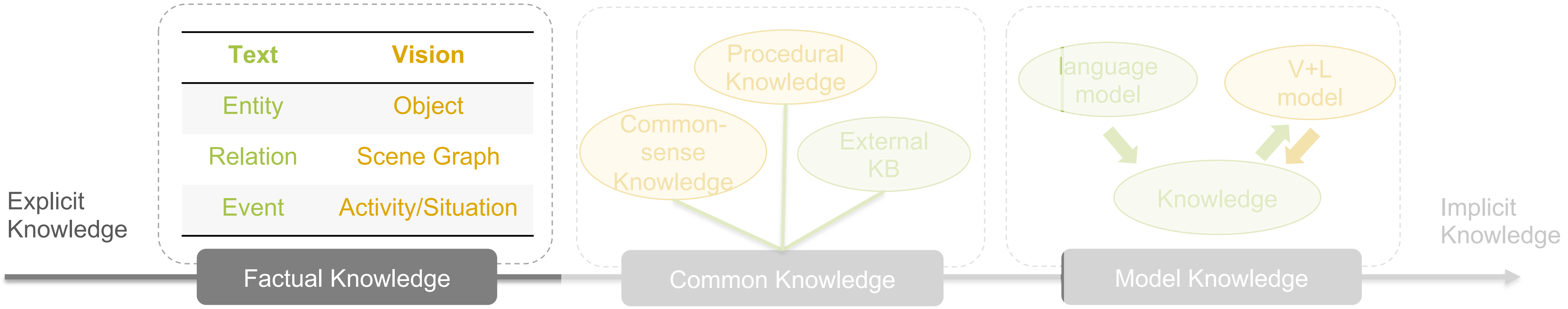


Effectiveness of Knowledge-Driven V+L Pretraining



Factual Knowledge

Compared to raw data, knowledge is **important and useful information.**



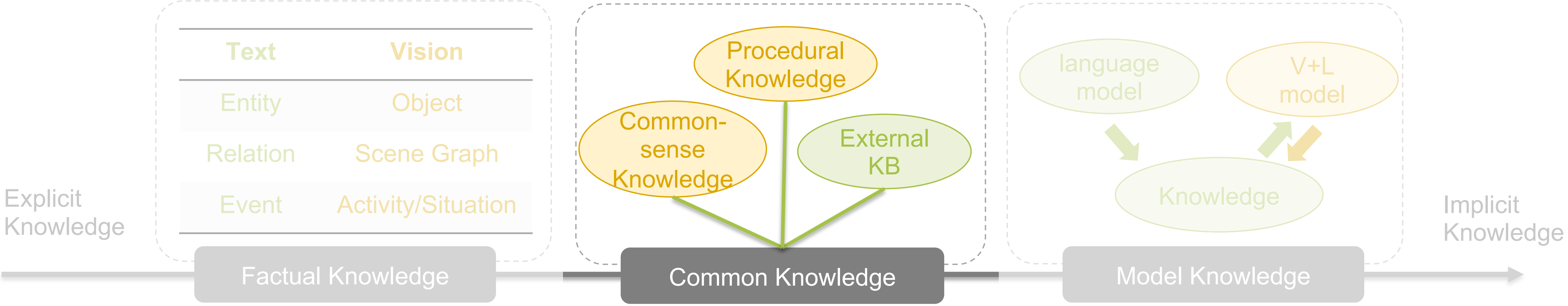
Text	**Vision**
Entity	Object
Relation	Scene Graph
Event	Activity/Situation

Factual Knowledge

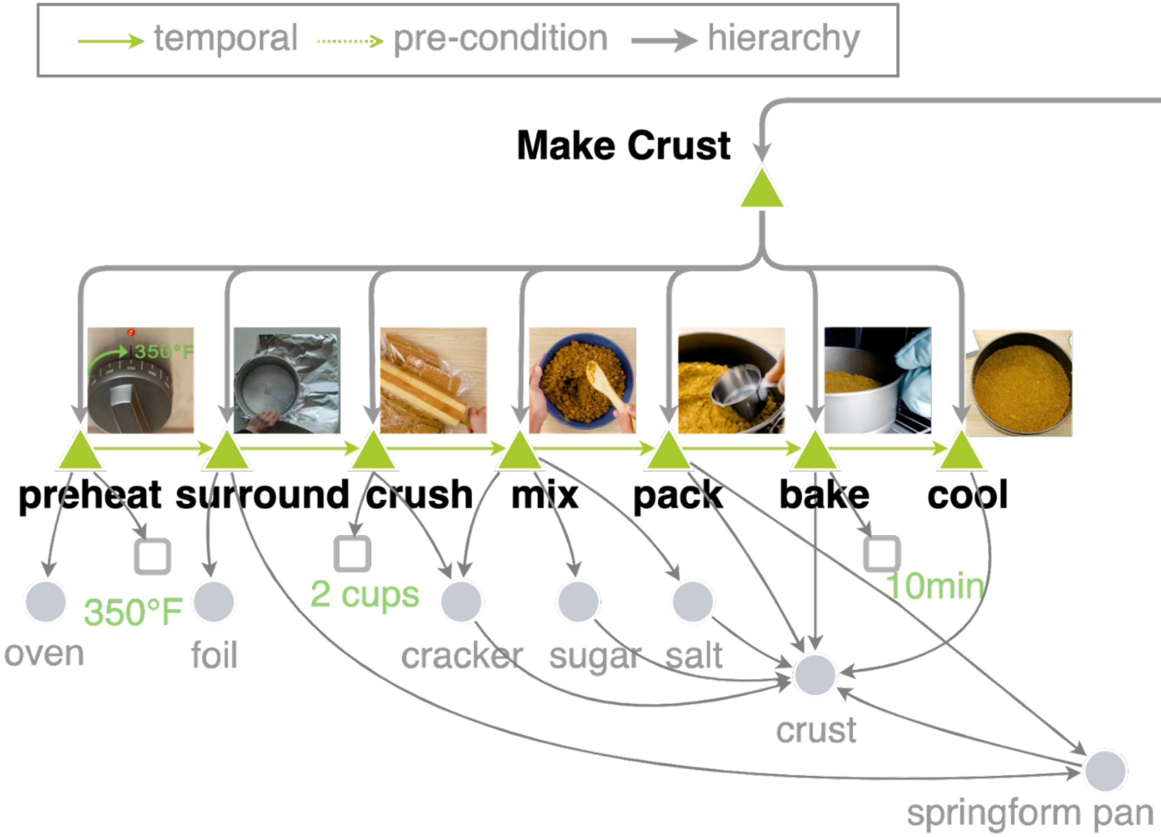
Entity	Hard Object Detection	Soft Object Prompt
SceneGraph	SceneGraph Parsing	Relational Masking
EventGraph	Activity-Based Models	Situation-Based Models

Common Knowledge

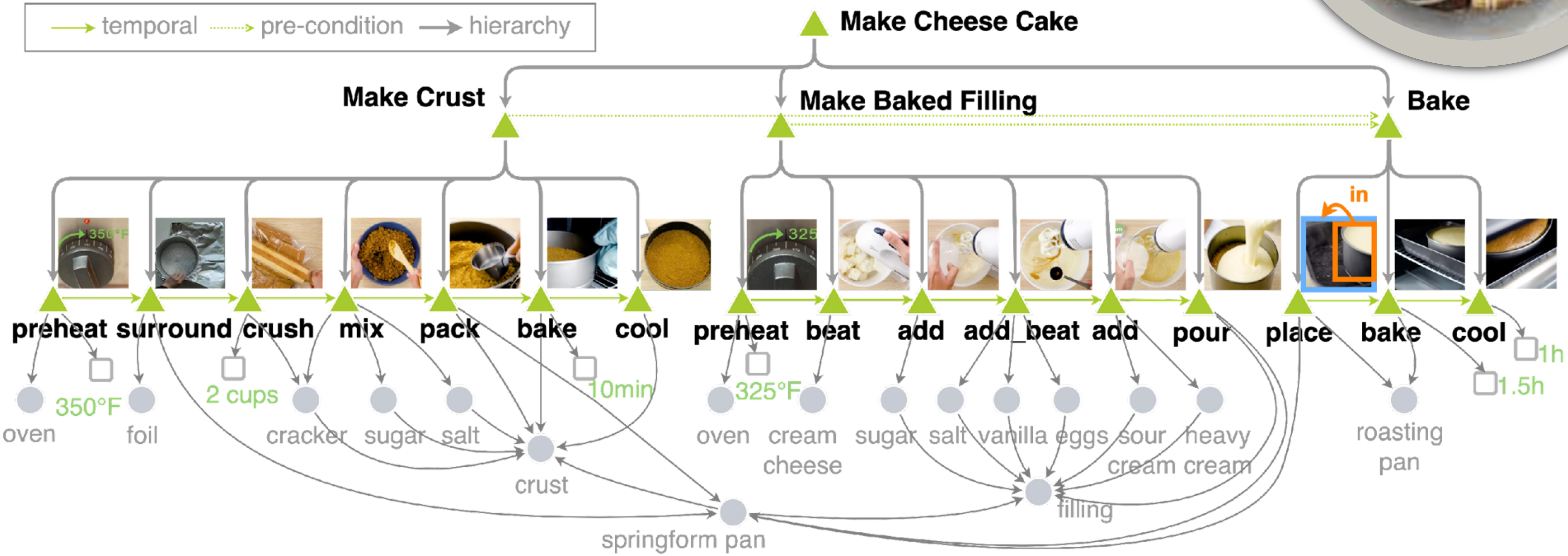
Compared to raw data, knowledge is **important and useful information.**



History repeats itself



History repeats itself



Procedural Knowledge

Representative Resource: wikiHow

Current online instructional video/text describe a task as completing a sequential set of steps, assuming that all tasks follow a linear schema

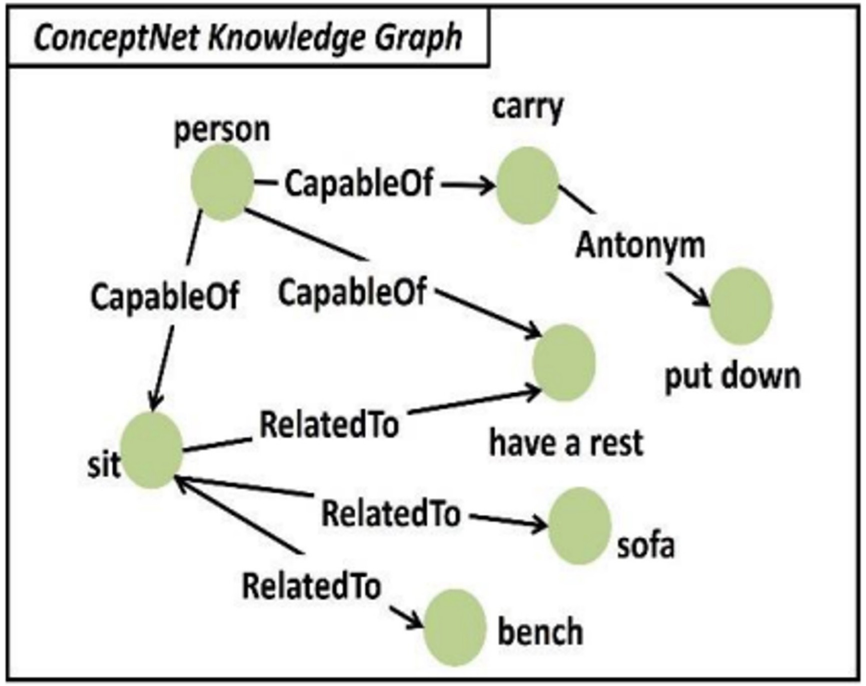
Heat Wok with Vegetable Oil → Put in Cold Rice and Cook 3min → Add Salt → Add light soy sauce → Put in Cooked & Salted Eggs → Add green onions → Serve

wikiHow to do anything...

YouTube

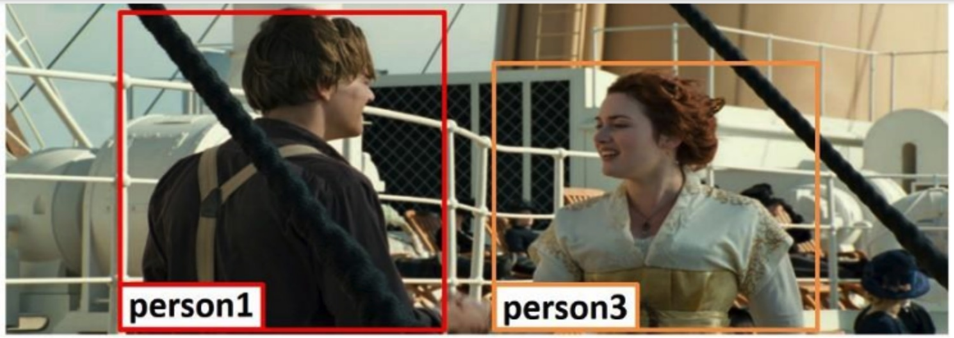
Commonsense Knowledge

Commonsense knowledge **includes facts about events occurring in time, about the effects of actions.**



ConceptNet
An open, multilingual knowledge graph

Freebase



Why are [person1] and [person3] shaking hands?

(a) [person1] and [person3] are presenting a trophy to someone.
(b) [person1] and [person3] just made a deal.
(c) [person1] and [person3] are old friends seeing each other for the first time in a long time.
(d) They have just met and are greeting each other.

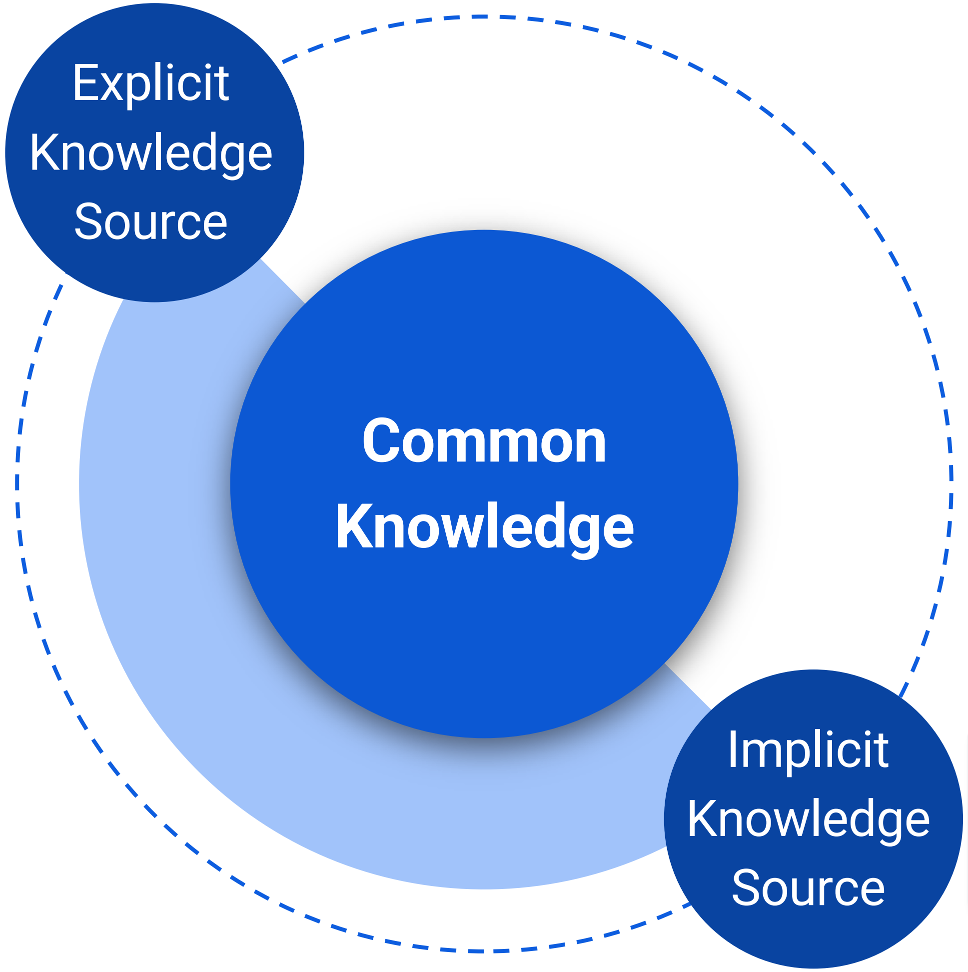
I think so because ...

(a) People like to greet each other when they meet by shaking hands.
(b) They look like they are shaking hands to introduce themselves.
(c) They are meeting each other for the first time.
(d) Some people shake hands to greet one another by grasping each others' arms.

How to add common knowledge?

- Two ways to learn procedural knowledge

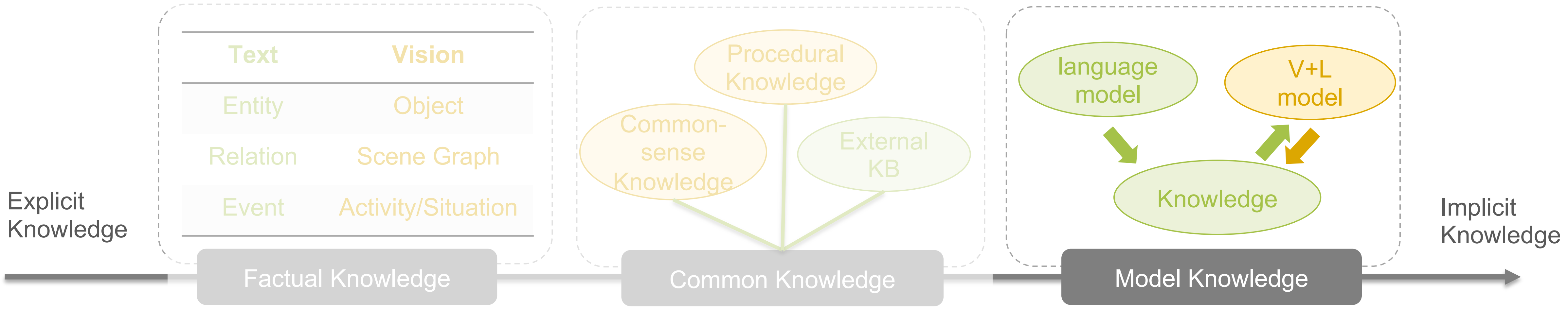
Use knowledge:
- As Data
- As Supervision
- In Model



Learning from massive data

Model Knowledge

Compared to raw data, knowledge is **important and useful information.**



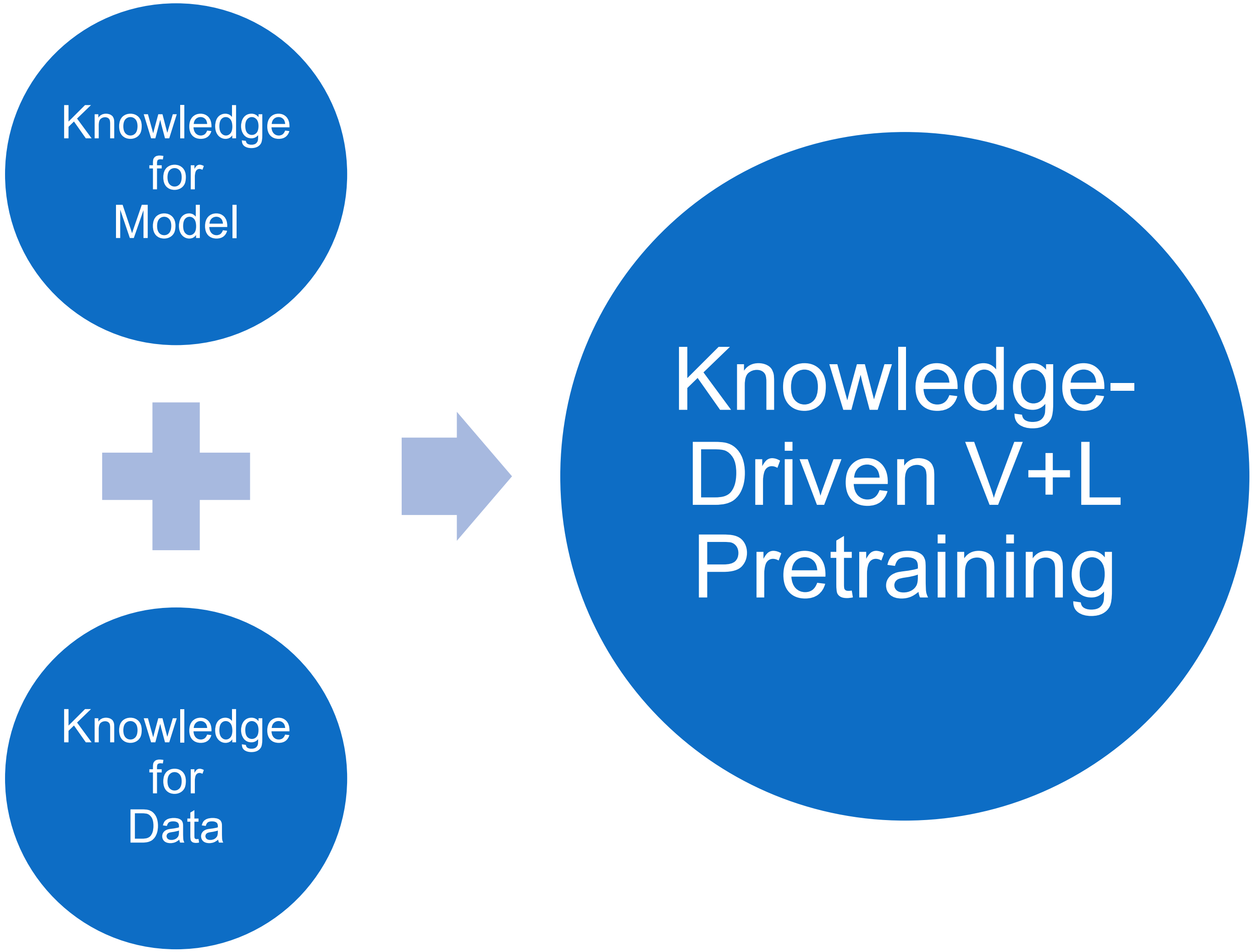
Model Knowledge

Implicit Knowledge → Leveraging Frozen LM → Convert multimodal info into text

Explicit Knowledge → External KB → Large Language Models

Knowledge Distillation → Physical Knowledge → Commonsense Knowledge

How to learn multimedia embedding?

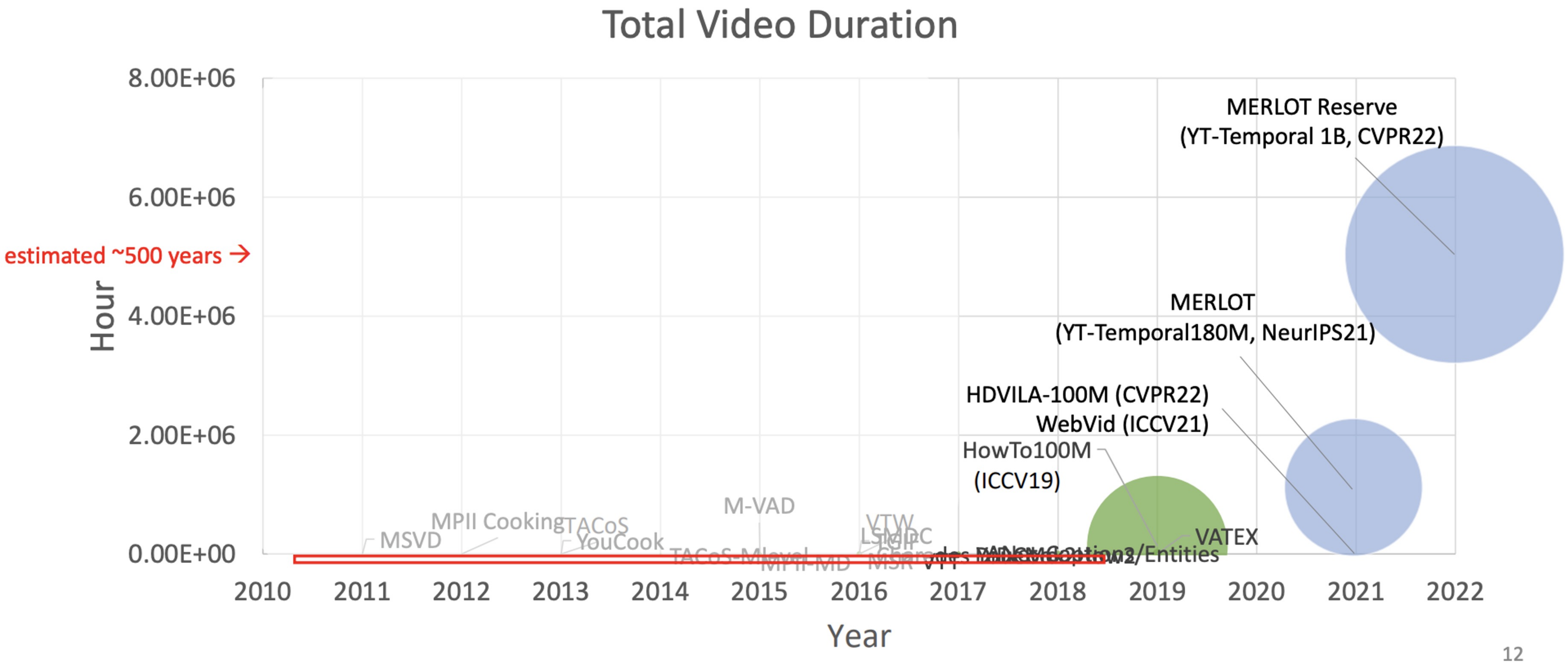


Resources: Image-and-Language Datasets

- Image-caption pairs
  - Free of cost
  - The state-of-the-art model CLIP (from OpenAI) uses ~400M pairs for model training
- Object and Scene Graph annotations in Popular datasets
  - Flicker 30K (~30K images)
  - MS COCO (~330K images)
  - Visual Genome (~108K images)
  - Conceptual Captions (~3.3M images)
  - SBU Captions (~1M images)
  - …

Learning Transferable Visual Models From Natural Language Supervision, OpenAI. 2021

Resources: Video-and-Language Datasets



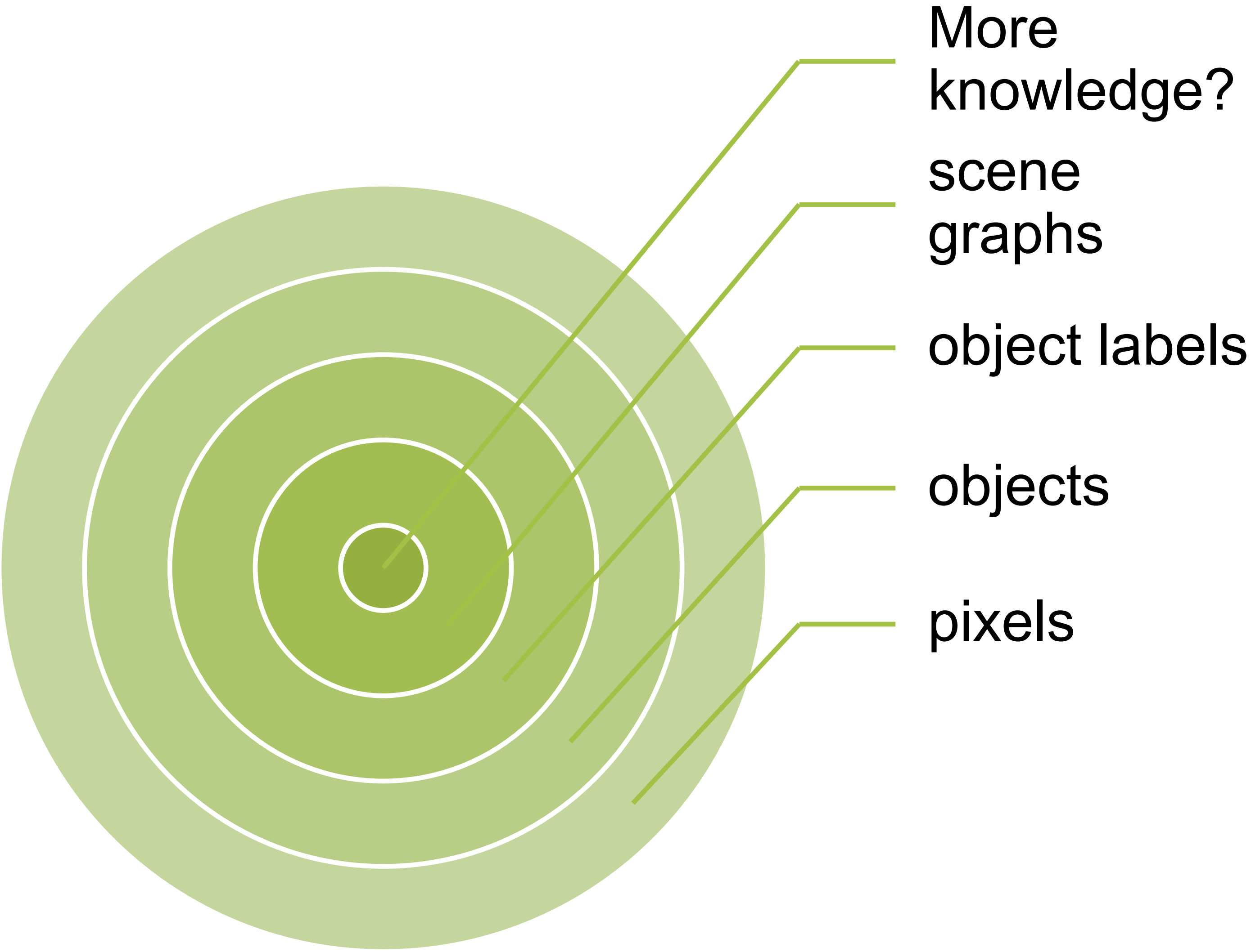
Merlot: Multimodal neural script knowledge models, NeurIPS 2021

https://datarelease.blob.core.windows.net/tutorial/VLP-Tutorial_2022/video_text_part1.pdf

Moving towards…

More knowledge

More data



Moving towards…

- On the model side, Transformer based models using self supervision has achieved great success in multiple downstream tasks.
  - Adding knowledge can guide the model where to focus.
  - Structured knowledge (such as event graph structure) and abstract word understanding (such as verb, adjectives, etc) are still lack of exploration.
- On the data side, knowledge is useful in the following ways:
  - In-context prompt
  - data augmentation
  - data selection

Timetable

Content	Time	Presenter
Motivation and Overview	15min	Manling Li
Factual Knowledge	15min	Manling Li
Procedural Knowledge	30min	Xudong Lin
Commonsense Knowledge and Model Knowledge	30min	Jie Lei
Panel: Knowledge vs Large Models	20min	Heng Ji, Mohit Bansal, Shih-Fu Chang
Panel: Text vs Image vs Video vs Others	20min	Heng Ji, Mohit Bansal, Shih-Fu Chang
Panel: Open Challenges	20min	Heng Ji, Mohit Bansal, Shih-Fu Chang
QA	30min	All