Spatial Commonsense Knowledge

Commonsense knowledge

Essential for many AI applications, including those in natural language processing, visual processing, and other tasks

- For instance, in natural language understanding and visual understanding it allows to cope with incomplete, ambiguous and noisy information

Humans learn commonsense knowledge from life events and experiences

Can we acquire and store commonsense knowledge so that the machine can use it in (spatial) language understanding?

Spatial common sense in knowledge bases

- Stored in knowledge bases like ConceptNet
- Here used in vision-language navigation task
- Utilize informative clues obtained from a KB for exploration

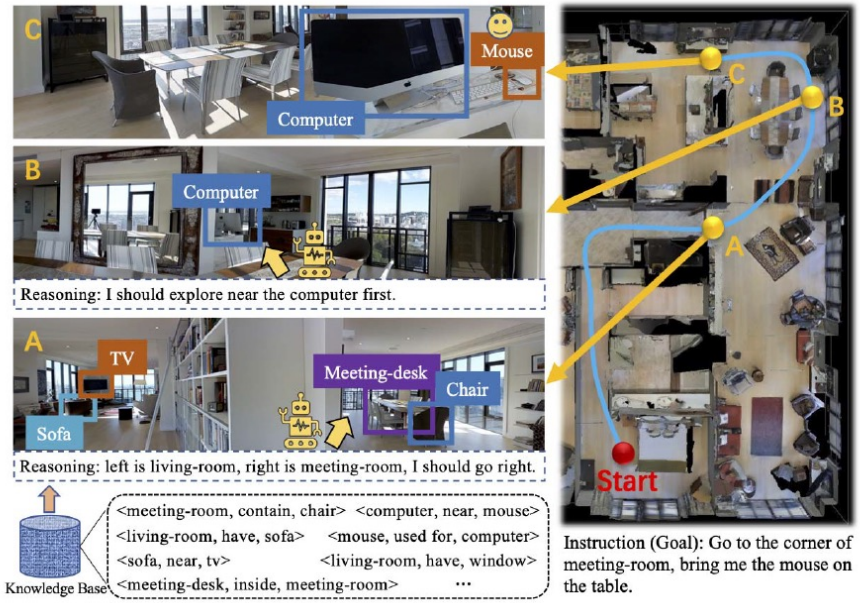


Figure 1. At viewpoint A, our agent with commonsense turns right into the 'meeting room' through perceived 'chair' and 'meeting-desk'. Then at viewpoint B, it seeks for easy-to-find related objects (*e.g.*, 'computer') at first for efficient exploration, where target 'mouse' is usually around. C is the final viewpoint it arrived.

Chen Gao, Jinyu Chen Si Liu, Luting Wang, Qiong Zhang, and Qi Wu (2021). Room-and-object aware knowledge reasoning for remote embodied referring expression. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition* (pp. 3064-3073).

Common sense in images paired with language

A girl rides a horse

From images paired with text

- = a spatial "question-answering" task where the question consists in a spatial commonsense query such as *where is the* "man" *located with respect to a* "horse" *when a* "man" *is* "feeding" *the* "horse"?
- The answer is a 2D "imagined" representation in contrast with a sentence/word as typically done in question-answering tasks
- How to learn this task:
  - Given a structured text input of the form (Subject, Relationship, Object) = (S,R,O)
  - Predict the 2D relative spatial arrangement of two objects (output)
- Train the task in a supervised setting:
  - Training set of image-text pairs, where the size and location of bounding boxes of objects in images serve as ground truth

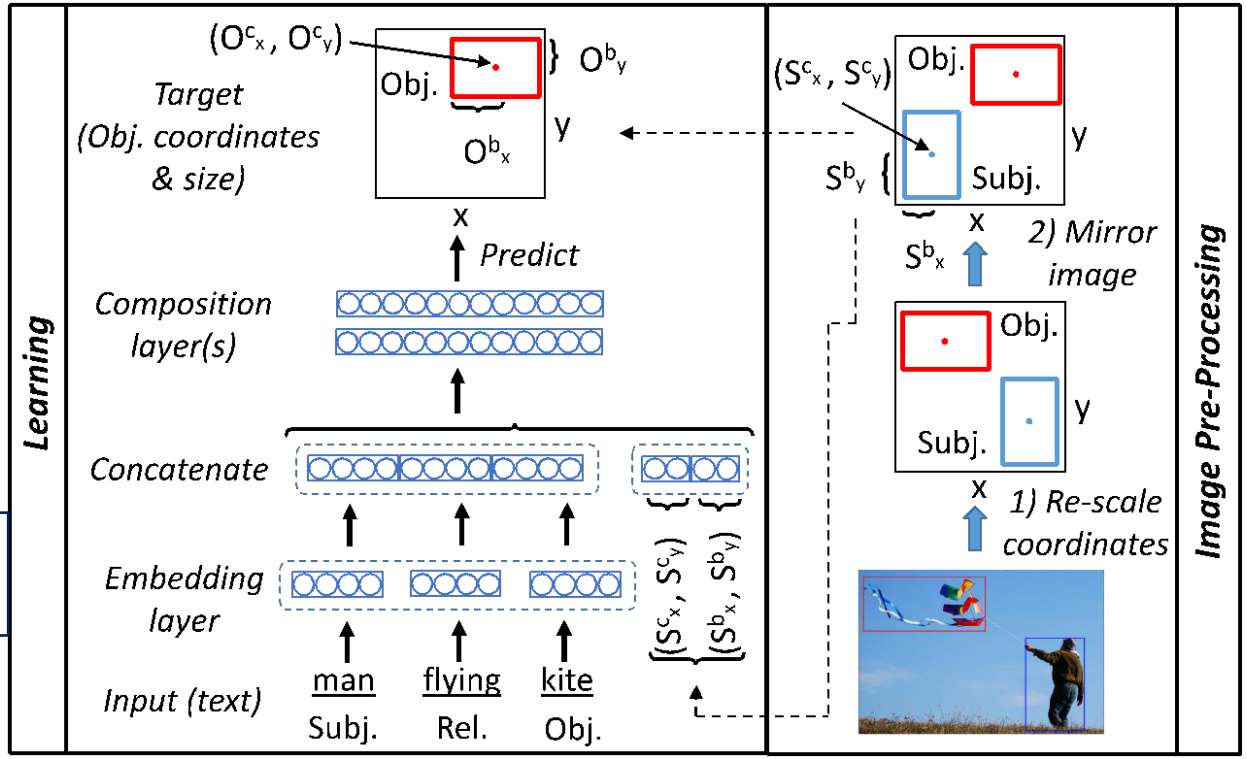
Guillem Collell and Marie-Francine Moens (2018). Learning representations specialized in spatial knowledge: leveraging language and vision. *Transactions of the Association for Computational Linguistics* (*TACL*), 6, 133-144.

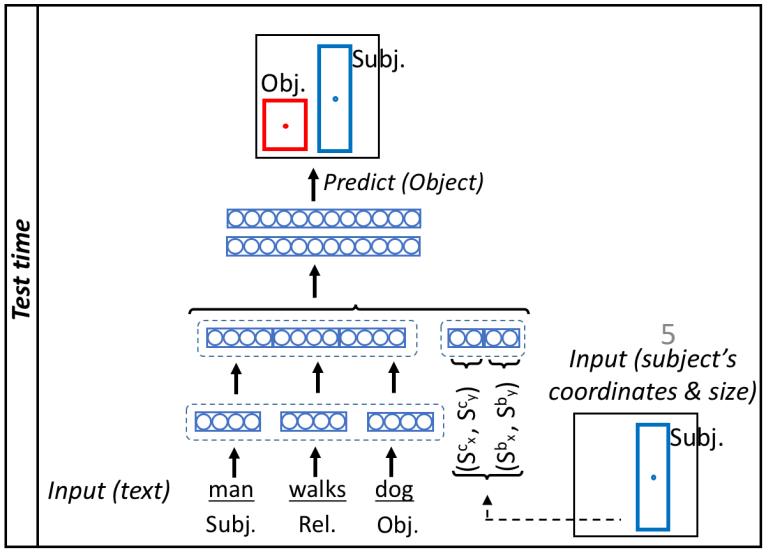
Simple feedforward neural network

Loss: mean squared error

Word embeddings to generalize over unseen words

Triplet of words, coordinates of subject





Guillem Collell, Luc Van Gool, and Marie-Francine Moens (2018). Acquiring common sense spatial knowledge through implicit spatial templates. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence* (AAAI 2018) (pp. 6765-6772). AAAI.

5

Common sense in images paired with language

Qualitative evaluation

Figure 2: Predictions by the model that leverages word embeddings (*EMB*). **Top**: Predictions in unseen words (underlined). **Bottom:** Predictions in unseen *triplets*.

Generalization through the word embeddings

Guillem Collell and Marie-Francine Moens (2017). Learning Visually Grounded Common Sense Spatial Knowledge for Implicit Spatial Language. In *Proceedings of the 13th International Symposium on Commonsense Reasoning*, University College London. CEUR.

Common sense in images paired with language

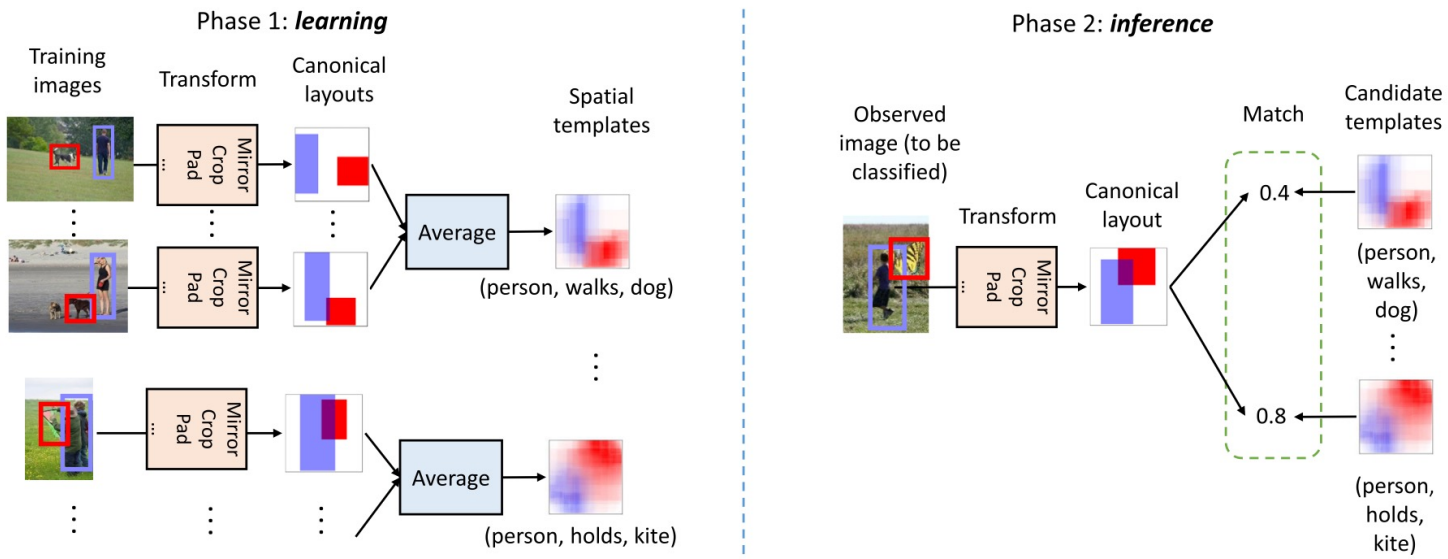


FIGURE 1. **General pipeline of scene understanding (human-object interaction) using our spatial template-based method. Spatial templates are first learned (left) and used afterwards to compute their semantic match with a given (observed) test image.**

Guillem Collell, Thierry Deruyttere, and Moens-Francine Moens (2021). Probing spatial cues: Canonical spatial templates for object relationship understanding. *IEEE Access 9*, 134298-134318.

Sizes and metric distances between objects

- Spatial reasoning tasks often involve the estimation of the size of an object or the metric distance between objects (e.g., when a robot or autonomous vehicles understands instructions in natural language)
- Valid sizes and distances are usually part of commonsense knowledge that humans possess when using language
- Sizes and distances could be estimated based on visual data

GPT-4V*

As an AI, I'm unable to physically interact with environments, but I can provide some insights based on the image provided. ... From the image, we can make some visual estimations. Judging from the image, it appears that the sofa and the table are quite close to each other, and there may not be enough space to accommodate a 1-meter wide robot....

Spatial-VLM

[VLM Reasoning] How wide is the path between the sofa and the table and chairs? [VLM Answer] 1.56m [Answer] Yes, the robot can go through the path between the sofa and the table and chairs since it is wider than the robot's width.

Boyuan Chen, Zhuo Xu, Sean Kirmani, Brian Ichter, Danny Driess, Pete Florence, Dorsa Sadigh, Leonidas Guibas, and Fei Xia (2024). SpatialVLM: Endowing vision-language models with spatial reasoning capabilities. arXiv: 2401.12168.

Sizes and metric distances between objects

- Given the importance of the topic, we witness a rise in visual question answering datasets that contain quantitative spatial question such as:

  "How much to the left is object X compared to object Y?"

  "How far is object X from object Y?"

- Large-scale spatial VQA dataset, SpatialVLM specifically designed for reasoning in quantitative metric spaces

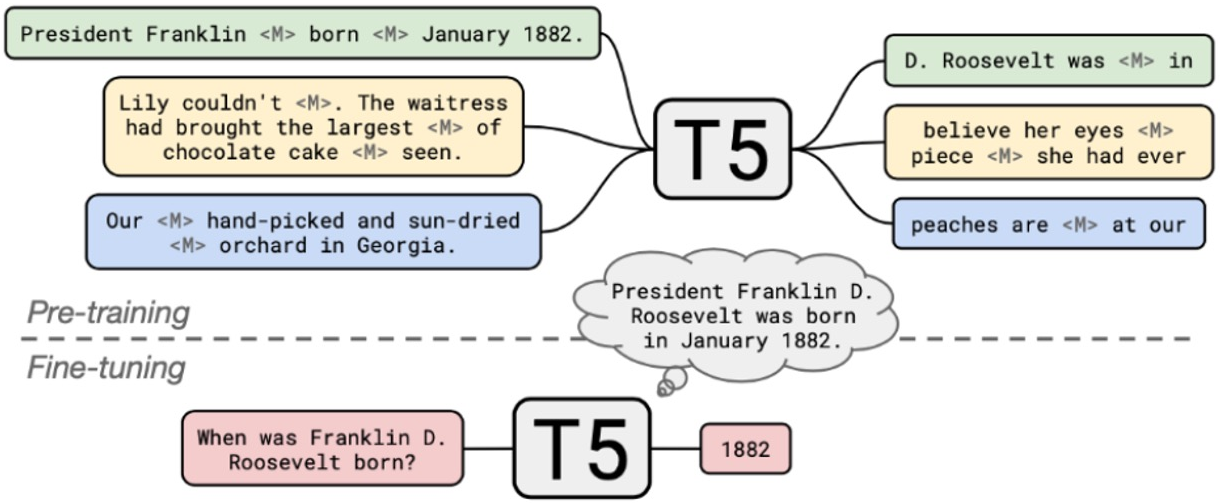
Sizes and metric distances between objects

- Distances can also be context-dependent

The distance between the man and the motorcycle is usually much smaller in a city environment compared to a highway environment

Common sense in language foundation models

- Pre-trained language model (e.g., trained on Wikipedia **texts**) act as a knowledge storage



Common sense in language foundation models

- LLMs contain knowledge about spatial arrangements of objects

  => useful for daily tasks of robots (e.g., setting a table, tidying up a room)

But spatial reasoning based on LLMs (e.g., translation by GPT-4 to a form that Wolfram Alpha can accept) is still difficult: especially world math problems that involved spatial reasoning

Zirui Zhao, Wee Sun Lee, and David Hsu (2023). Large language models as commonsense knowledge for large-scale task planning, 2023. *In Proceedings of the NeurIPS 2023 Workshop on Foundation Models for Decision Making*.

Ernest Davis (2024). Mathematics, word problems, common sense, and artificial intelligence. *Bulletin of the American Mathematical Society*.

Common Sense in Visual and V&L Foundation Models

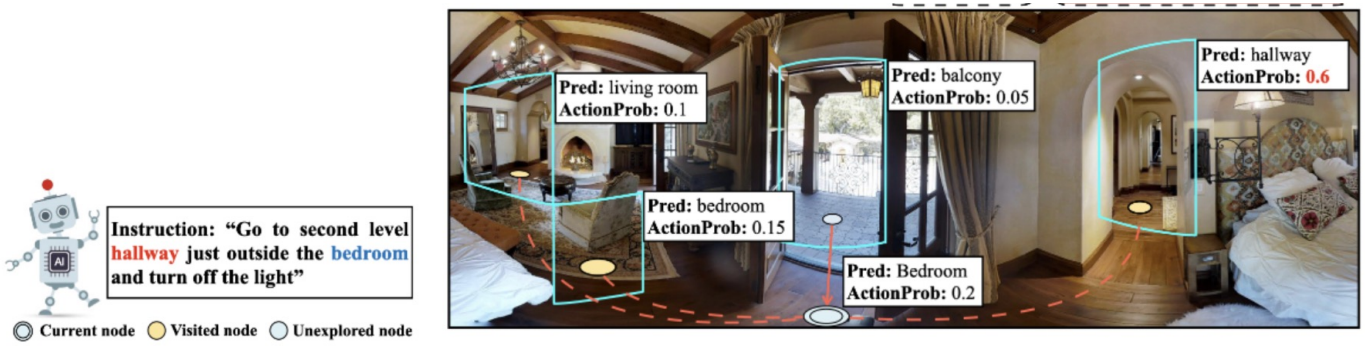
- Especially V&L foundation models (e.g., CLIP): remarkable performance in visual question answering
- Emerging **video foundation models** provide commonsense knowledge with respect to postconditions of actions and resulting object locations

Boyuan Chen, Zhuo Xu, Sean Kirmani, Brian Ichter, Danny Driess, Pete Florence, Dorsa Sadigh, Leonidas Guibas, and Fei Xia (2024). SpatialVLM: Endowing vision-language models with spatial reasoning capabilities. arXiv:2401.12168

Rowan Zellers, Jiasen Lu, Ximing Lu, Youngjae Yu, Yanpeng Zhao, Mohammadreza, Salehi, Aditya Kusupati, Jack Hessel, Ali Farhadi, and Yejin Choi (2022). Merlotreserve: Neural script knowledge through vision and language and sound. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (CVPR) (pp. 16375–16387).

Yi Wang et al. (2024). InternVideo2: Scaling video foundation models for multimodal video understanding. arXiv:2403.15377

Spatial common sense in a navigation task



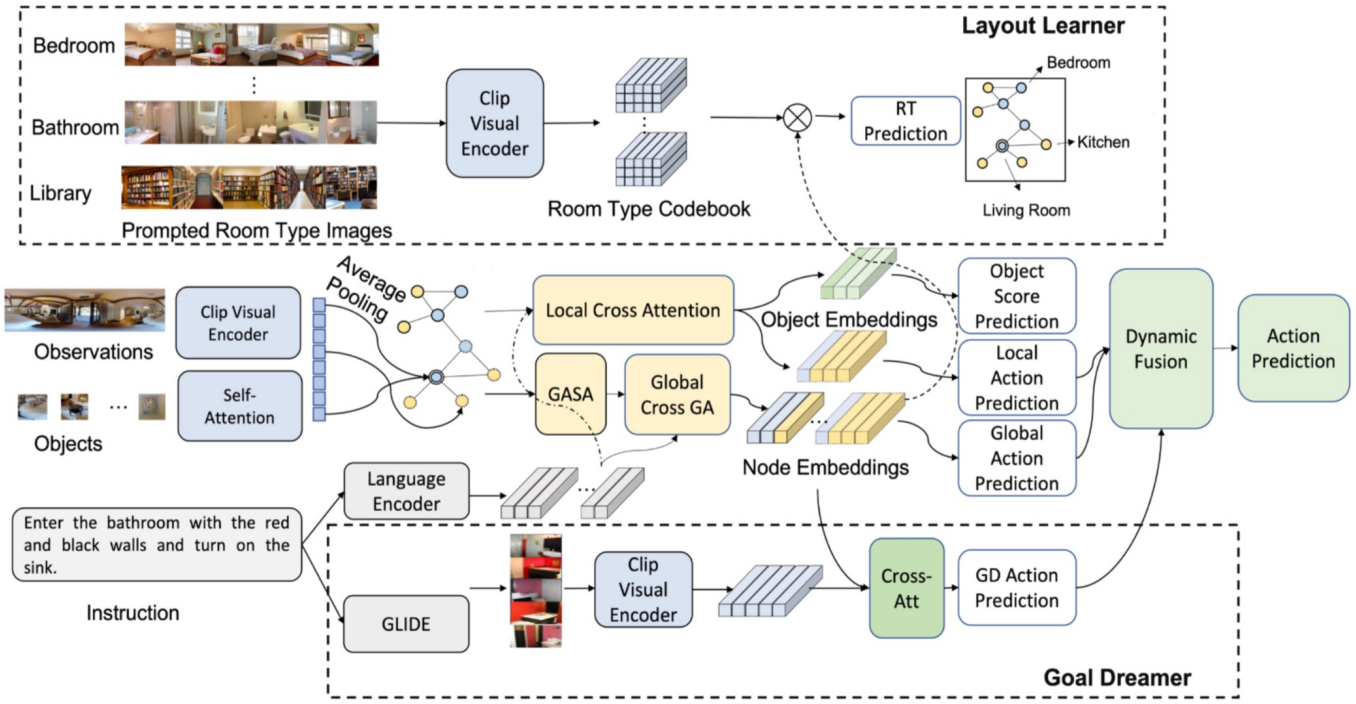
- Task setup: Given a high-level instruction where the agent needs to find the object described by the instruction
  - As the agent could be anywhere in the envrionment, to correctly find the object, it needs to navigate to the position where the object becomes visible, then the agent can identiy the correct objects among all visible objects
- Challenges: How to effectively explore the environment ?

  How to better generalize to previously unseen environment ?

Mingxiao Li*, Zehao Wang*, Tinne Tuytelaars, and Marie-Francine Moens. Layout-aware dreamer for embodied referring expression grounding (2022). In *Proceedings of the AAAI Conference on Artificial Intelligence* 37 (1), 1386-1395.

Spatial common sense in a navigation task

Learns to infer the room category distribution of neighboring unexplored areas along the path → effectively introduces **layout common sense of room-to-room transitions**



Baseline model: VLN-DUET [1]

Leverage the knowledge in pretrained text-image generation model → imagine the destination beforehand to help the agent to conduct more effective exploration

[1] Chen, S., Guhur, P.-L., Tapaswi, M., Schmid, C. & Laptev, I. 2022. Think Global, Act Local: Dual-scale Graph Transformer for Vision-and-Language Navigation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 16537–16547.

Mingxiao Li*, Zehao Wang*, Tinne Tuytelaars, and Marie-Francine Moens. Layout-aware dreamer for embodied referring expression grounding (2022). In *Proceedings of the AAAI Conference on Artificial Intelligence* 37 (1), 1386-1395.

Cases where common sense is insufficient

- GPT-4, a large multi-modal foundation model, when given two facts: R1(x,y) and R2(y,z): what RCC-8 relations are possible between x and z?

- GPT-4 was able to fill in 70% of the composition table
  - Sometimes relation was confused with its inverse (maybe due to language encoding that ignores spatial structure?)

Anthony G. Cohn (2023). An evaluation of ChatGPT-4's qualitative spatial reasoning capabilities in RCC-8. arXiv 2309.15577.

Cases where common sense is insufficient

- Foundation models struggle with long-tail knowledge
- LLMs when used in object layout planning fail to generate suitable layouts for objects that involved in unusual or unexpected spatial relationships
  - Fast and slow models? Former based on common sense in foundation models, the latter based on object identification and reasoning?

Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel (2023). Large language models struggle to learn long-tail knowledge. In *Proceedings of the 40th International Conference on Machine Learning* (ICML 2023).

Sherry Yang, Ofir Nachum, Yilun Du, Jason Wei, P. Abbeel, and Dale Schuurmans (2023). Foundation models for decision making: Problems, methods, and opportunities. *ArXiv* 2303.04129.

Ruben Cartuyvels, Wolf Nuyts, and Marie-Francine Moens (2024). Explicitly representing syntax improves sentence-to-layout prediction of unexpected situations. *Transactions of the Association for Computational Linguistics*, 12: 264–282, 2024.

Spatial Language Grounding and Translation of Spatial Language to Coordinates in a 2D or 3D Physical World

- It is well-known that humans "imagine" language content in a visual space
- It is well-known that humans reason in spatial visual space
- Even if we reason with symbolic representations, many practical applications require identifying the location in the physical world where events/actions have happened, they will happen or need to happen

- How to predict the spatial configurations and location of objects, actions, and their attributes in a 2D or 3D space?

Language grounding in 2D or 3D physical space

- This work has potential for real-time language understanding in a visual context:
  - Language communication to robots, machines, self-driving cars, …
  - Translation of spatial language to 2D or 3D space opens possibilities of **quantitative reasoning in such a space**, which can complement qualitative symbolic representations and reasoning

- This work is a step towards evaluating spatial language understanding by visualizing the interpreted content

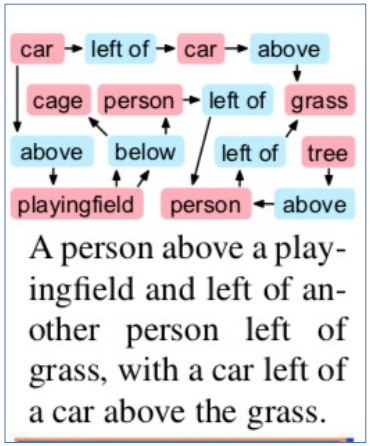
Text-to-image/video synthesis

Text-to-image and text-to-video synthesis techniques aimed at naturally composing and visualizing **text** instances:

- Many practical applications in education, gaming, creating virtual realities steered and manipulated through language

T2I synthesis: integration of a scene graph

- Text is first translated into a scene graph (= symbolic representation expressing the objects and their semantic/spatial relationships)
- The spatial layout is generated from the scene graph



- Use of a graph convolution network composed of several graph convolution layers to represent objects and their relationships
- Followed by steps of layout prediction and pixel prediction

T2I synthesis: integration of a scene graph

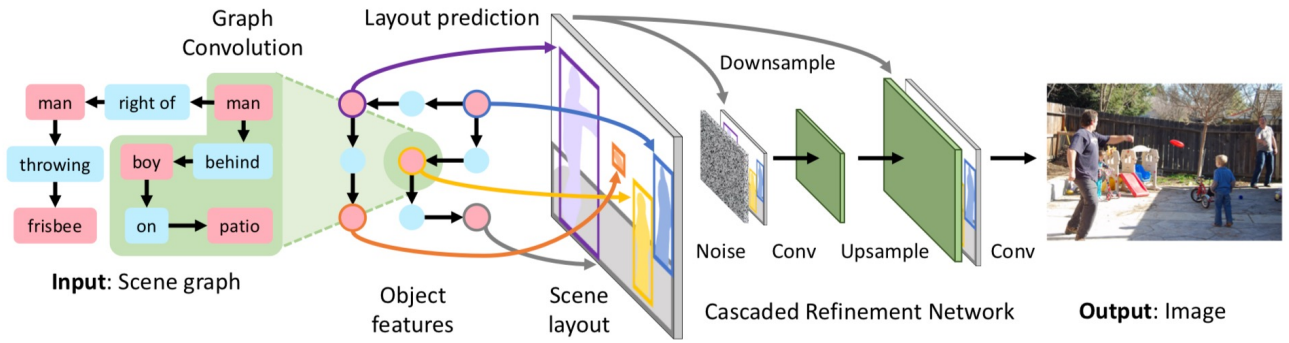


Figure 2. Overview of our image generation network f for generating images from scene graphs. The input to the model is a *scene graph* specifying objects and relationships; it is processed with a *graph convolution network* (Figure 3) which passes information along edges to compute embedding vectors for all objects. These vectors are used to predict bounding boxes and segmentation masks for objects, which are combined to form a *scene layout* (Figure 4). The layout is converted to an image using a *cascaded refinement network* (CRN) [6]. The model is trained adversarially against a pair of *discriminator networks*. During training the model observes ground-truth object bounding boxes and (optionally) segmentation masks, but these are predicted by the model at test-time.

Justin Johnson, Agrim Gupta, and Li Fei-Fei (2018). Image generation from scene graphs. In *Proceedings of the Conference on Computer Vision and Pattern Recognition* (CVPR).

T2I synthesis: integration of a scene graph

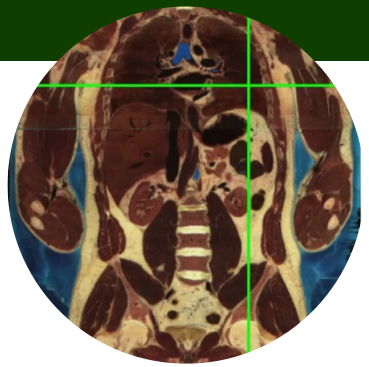
More details

- Scene graphs were manually created
- Many semantic relationships are spatial
- A generative adversarial network was trained end-to-end including several loss functions
- Interesting to mention is the box loss for layout prediction:
  - *Box loss:* $\mathcal{L}_{box} = \sum_{i=1}^{n} \left\| b_i - \hat{b}_i \right\|_1$ which penalizes the L_1 difference between ground-truth b_i and predicted box $\hat{b}_i$, where $n =$ number of objects in the graph
  - Optimized over all N training data

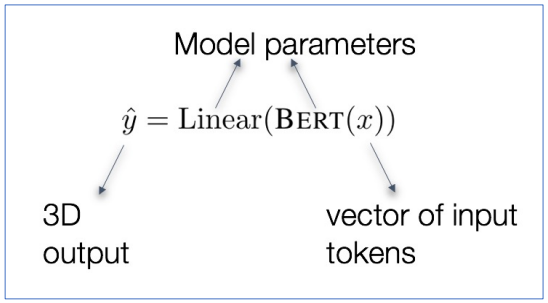
- Problem of semantic standards for object and relationship names in the scene graph

Justin Johnson, Agrim Gupta. and Li Fei-Fei (2018). Image generation from scene graphs. In *Proceedings of the Conference on Computer Vision and Pattern Recognition* (CVPR).

Map text to 3D



- BERT backbone
- Model input ⟝ Medical text tokenized with WordPiece
- Model output ⟝ [CLS] token representation projected into 3D



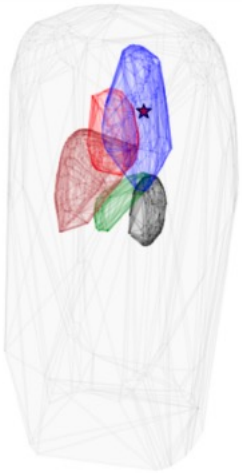


Figure 1: Given the text with implicit reference to lungs: "Divided into two lobes, an upper and a lower lobe, by the oblique fissure, which extends from the costal to the mediastinal surface" (Drake et al., 2009), our model learns the grounding indicated by the star.

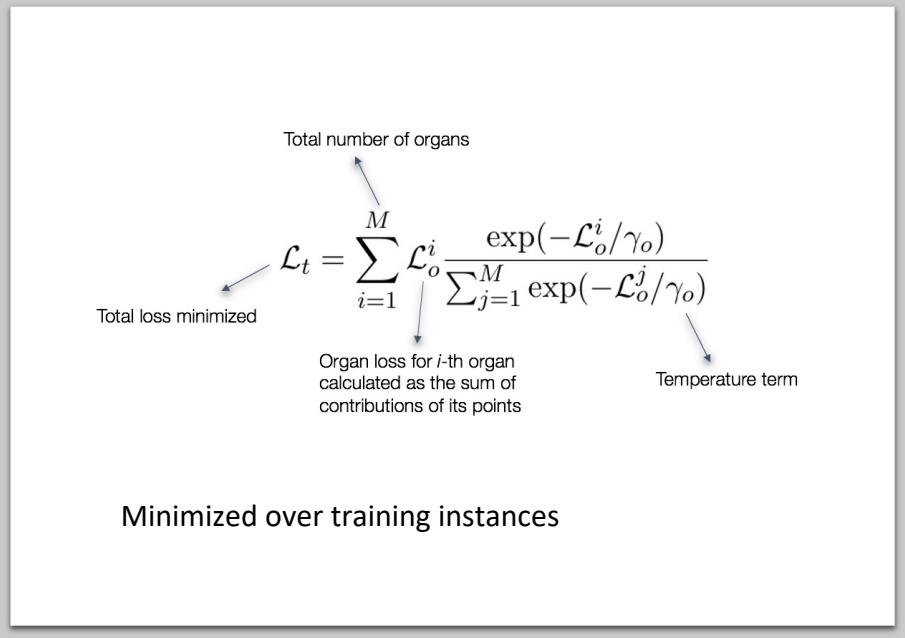
- Loss function: Enables **reasoning about** the semantic relatedness of medical text

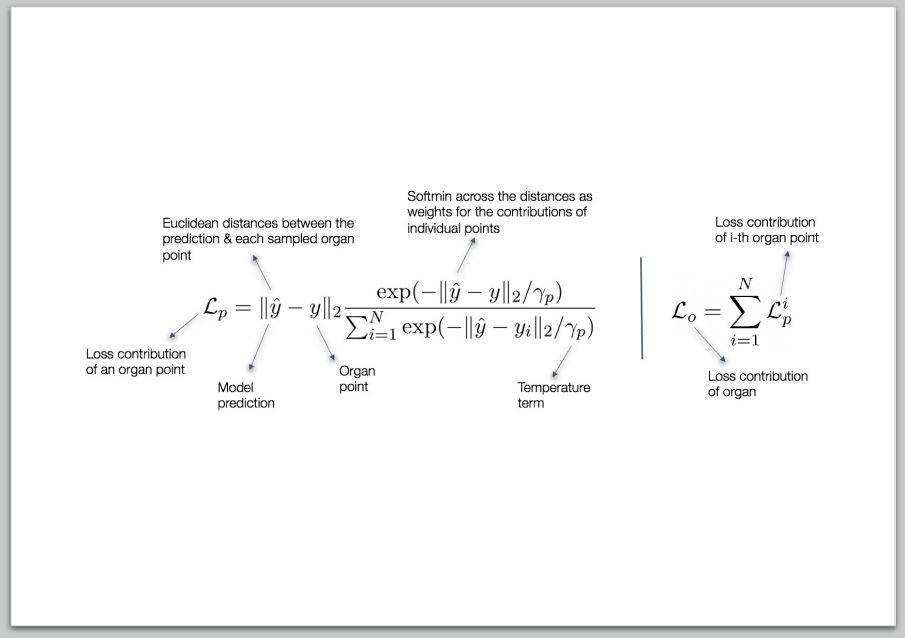
Dusan Grujicic, Gorjan Radevski, Tinne Tuytelaars, and Matthew Blaschko (2020). Learning to ground medical text in a 3D human atlas. In *Proceedings of the Conference on Computational Natural Language Learning* (CoNLL). ACL.

Map text to 3D

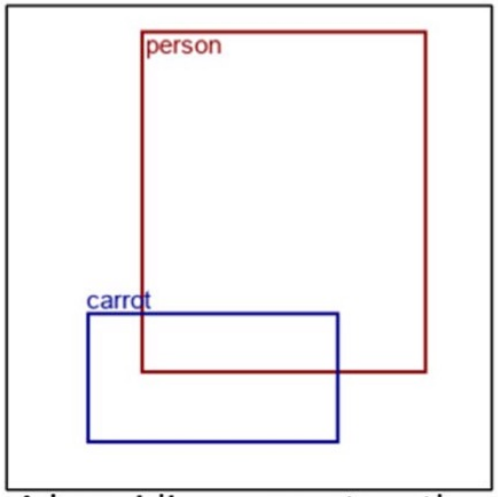
Soft Organ Distance loss

- Not only grounding the medical article to the right organ but also to the appropriate location within the organ based on the other organs mentioned as context without any explicit annotations at that level of granularity
- Could be refined by considering spatial language

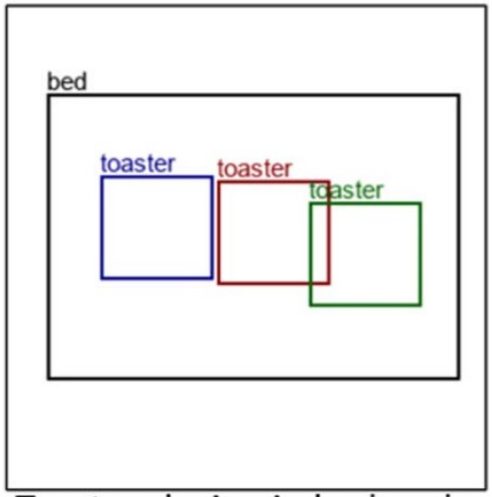




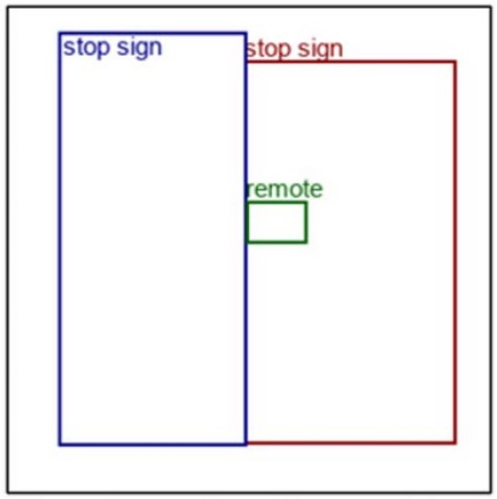
Dusan Grujicic, Gorjan Radevski, Tinne Tuytelaars, and Matthew Blaschko (2020). Learning to ground medical text in a 3D human atlas. In *Proceedings of the Conference on Computational Natural Language Learning* (CoNLL). ACL.



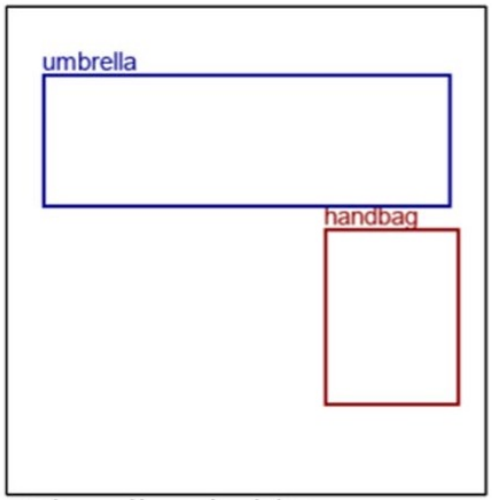
A boy riding a carrot on the ocean waves.



Toasters laying in bed under the covers.



Stop signs standing next to each other with one holding a video game controller.



A handbag holding an umbrella with a British flag design.

Emposing a structural loss in layout prediction

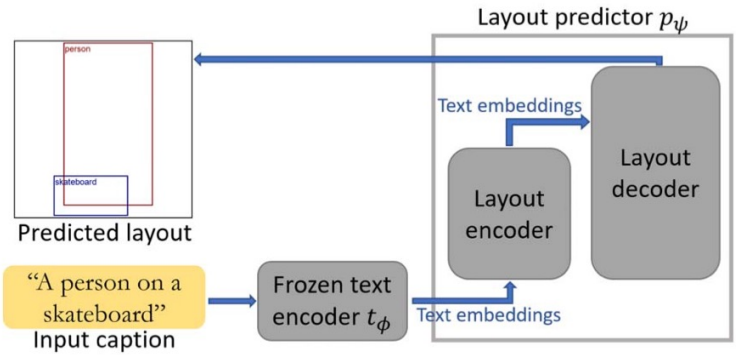


Figure 2: Overview of text-to-layout prediction.

- Constrastive structural loss enforces:
  - The grammatical structure found in the parse tree of the input sentence into the representations used by the layout predictor
  - That the object representations to be close to the tree positional embeddings of the sentence, but far from tree positional embeddings of other sentences

Ruben Cartuyvels, Wolf Nuyts, and Marie-Francine Moens (2024). Explicitly representing syntax improves sentence-to-layout prediction of unexpected situations. *Transactions of the Association for Computational Linguistics*, 12: 264–282, 2024.

Spatial reasoning for robot manipulation

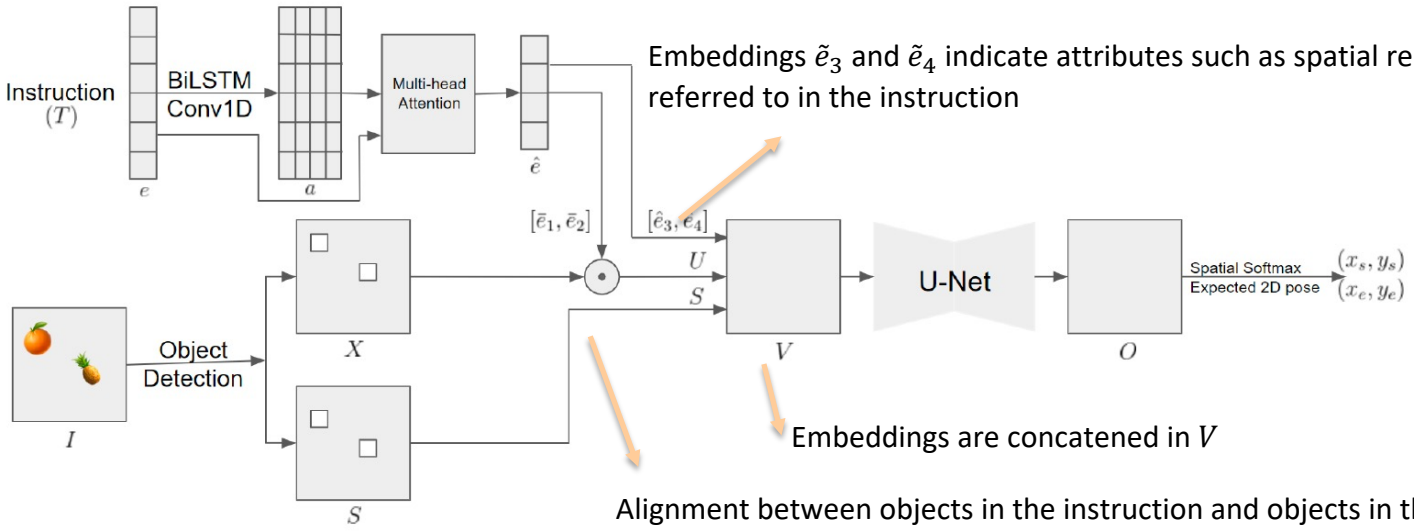


Fig. 3. The Lang-UNet model takes the instruction text T and the detected objects X along with their sizes S to predict the *start* and *end* co-ordinates. $\odot$ represents pixel-wise correlation. The U-Net has four Conv5x5(128)-ELU layers followed by four transposed convolutions of the same size and a bottleneck Conv1x1(2) layer. Section IV explains the network in more detail.

Compositional generalization of the objects and their spatial relations:

- The U-Net (convolutional hourglass network) has no notion of which object is at a specific position; it is only aware that a particular object selected via $\tilde{e}_1$ or $\tilde{e}_2$ is present
- So, U-NET learns embeddings of spatial relationships which allow to predict start and end coordinates: if the model has learned to find the position of *an apple to the left of the banana*, it will generalize to *an orange left to the banana*

Sagar Gubbi Venkatesh, Anirban Biswas, Raviteja Upadrashta, Vikram Srinivasan, Partha Talukdar, and Bharadwaj Amrutur (2021). Spatial reasoning from natural language instructions for robot manipulation. In *Proceedings of the IEEE International Conference on Robotics and Automation* (ICRA 2021).

From qualitative to quantitative reasoning

- Suppose qualitative spatial relationships available in annotated text data to train a model, how can we leverage these in a model that predicts the relative position of the objects on a 2D canvas ?
- Model could be easily extended to a prediction in a 3D physical space

From qualitative to quantitative reasoning

- Suppose a robot that must place objects on a 2D canvas following natural language instructions that describe the relative position between two objects (e.g., the book is to the left of the computer, a man rides a horse)
- The relative position of an object o_i is defined by its x and y coordinates on the 2D canvas
- Suppose that the objects were already detected in the instruction, but we do not have an annotated training dataset that pairs an object with its 2D coordinates of the canvas
- However, we do have a training set of instructions that are annotated with the spatial relations *left*, *right*, *under*, and *above* between each two objects mentioned (from the viewpoint of the robot)
- Performing spatial reasoning, we can reduce the relations to *right* and *above*

From qualitative to quantitative reasoning

- A spatial triplet (r_{ij}, o_i, o_j) is composed of a spatial predicate r_{ij} and its object arguments o_i and o_j
- When we train a model that predicts the relative positions of the objects in the 2D physical space, a common loss might compute the negative log likelihood of the triplet and minimize it
- And combined (e.g., in a sum) with losses that promote the relative positioning of the objects:

$$if \quad r_{ij} = right, \quad x_i > x_j$$
$$if \quad r_{ij} = above, \quad y_i > y_j$$

- Resulting in the following losses:

$$L_{right} = max(0, \hat{x}_j + m - \hat{x}_i)$$
$$L_{above} = max(0, \hat{y}_j + m - \hat{y}_i)$$

where m is a margin

E.g., the predicted point $\hat{x}_j$ should be at least a distance m smaller than predicted point $\hat{x}_i$ for this loss to be zero

The proposed losses are summed over the N training examples

- Current diffusion models (e.g., DALLE-E) work on top of CLIP representations and generate full scenes conditioned on an input text prompt
- However, they lack spatial reasoning skills to determine the correct positioning of objects given in the input (e.g., a suitcase is left to the person)

Leigang Qu, Shengqiong Wu, Hao Fei, Liqiang Nie, and Tat-Seng Chua (2023). LayoutLLM-T2I: Eliciting layout guidance from LLM for text-to-image generation.

Ruben Cartuyvels, Wolf Nuyts, and Marie-Francine Moens (2024). Explicitly representing syntax improves sentence-to-layout prediction of unexpected situations. Transactions of the Association for Computational Linguistics, 12: 264–282, 2024.

Giving a command to your self-driving car

C4AV @ ECCV 2020

COMMANDS FOR AUTONOMOUS VEHICLES WORKSHOP
23 AUGUST 2020 – GLASGOW

Challenge

- The task of visual grounding requires locating the most relevant region or object in an image, given a natural language query

(a) You can park up ahead behind **the silver car**, next to that lamppost with the orange sign on it

(b) **My friend** is getting out of the car. That means we arrived at our destination! Stop and let me out too!

(c) Yeah that would be **my son** on the stairs next to the bus. Pick him up please

(d) After **that man in the blue top** has passed, turn left

(e) There's **my mum**, on the right! The one walking closest to us. Park near **her**, she might want a lift

(f) Turn around and park in front of **that vehicle in the shade**

Fig. 2: Some examples from the Talk2Car dataset [5]. Each command describes an action that the car has to execute relevant to a referred object found in the scene (here indicated by the red 3D-bounding box). The referred object is indicated in bold in each command. Best seen in color.

Thierry Deruyttere, Simon Vandenhende, Dusan Grujicic, Yu Liu, Luc Van Gool, Matthew Blaschko, Tinne Tuytelaars & Marie-Francine Moens (2020). Commands 4 autonomous vehicles (C4AV) workshop summary. In *Proceedings of the 16th European Conference on Computer Vision.*

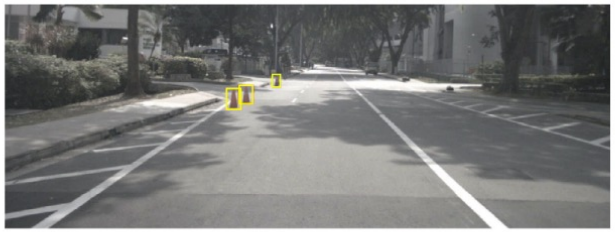
Spatial dialogue to resolve ambiguity

- A component for visual uncertainty analysis of the referred objects

- How can we report back the uncertainty of the self-driving to the passenger? => By generating sentences based on predicted attributes taking into account the features of uncertain objects

(a) The objects that cause uncertainty for the command: "Parallel park behind the car on the left".

(b) The objects that cause uncertainty for the command: "Change lanes and get behind the white car".

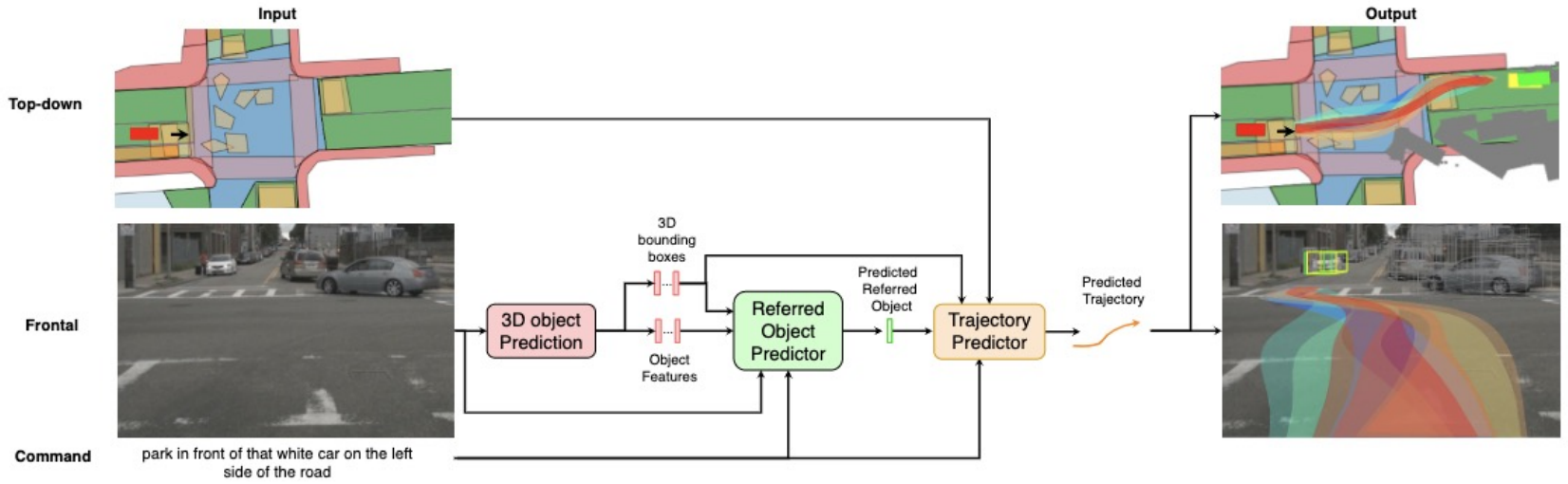


(c) The objects that cause uncertainty for the command: "After that signaling cone, turn left".

Fig. 5. Uncertainty Examples. Examples of uncertain objects detected in different scenes. We see that the objects flagged as uncertain by URS are often from the same (super)class. Best viewed in color. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Thierry Deruyttere, Victor Milewski, and Marie-Francine Moens (2021). Giving commands to a self-driving car: How to deal with uncertain situations? *Engineering Applications of Artificial Intelligence*, 103, Art. No. 104257, 1-20.

Destination and trajectory prediction



Grujicic, Dusan, Deruyttere, Thierry, Moens, Marie-Francine, and Blaschko, Matthew B. (2022). Predicting physical world destinations for commands given to self-driving cars. In *Proceedings of the Thirty-Sixth AAAI Conference on Artificial Intelligence* (AAAI 2022) (pp. 715-725). AAAI.

Temporal Language Grounding and Translation of Temporal Language to 1D Timeline

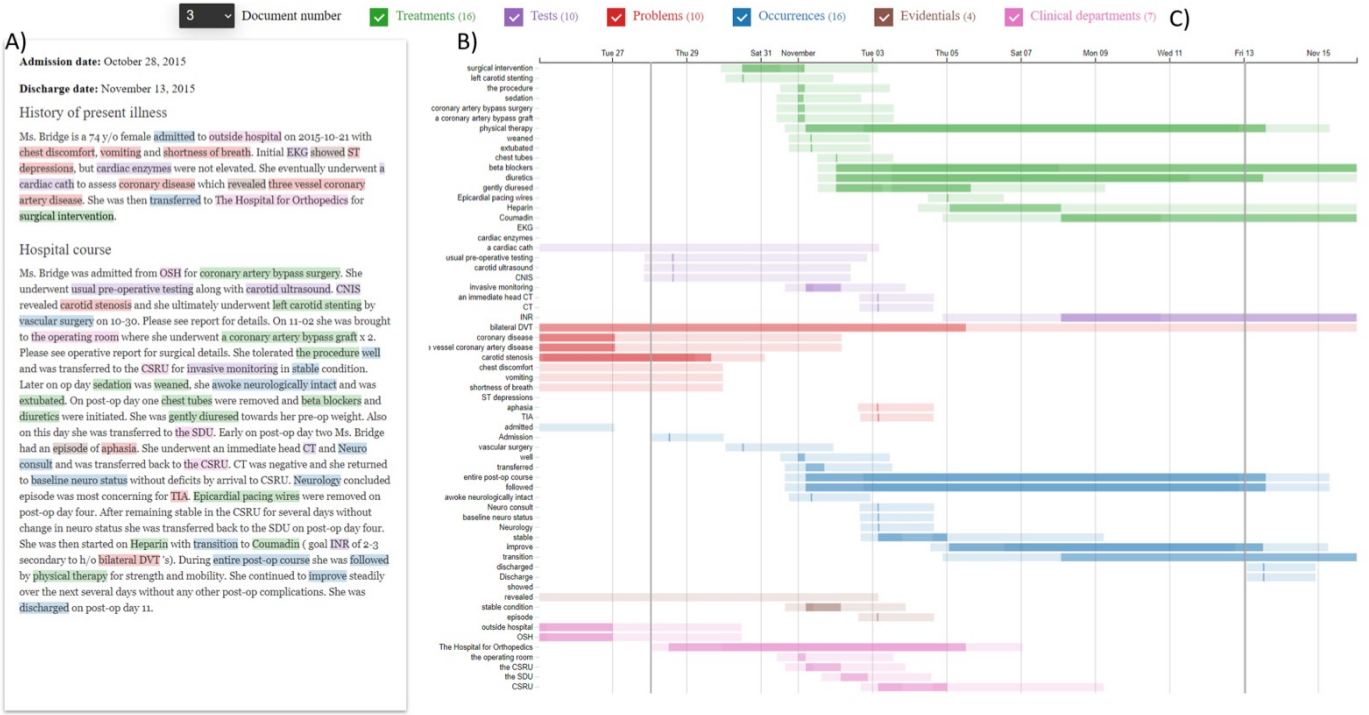
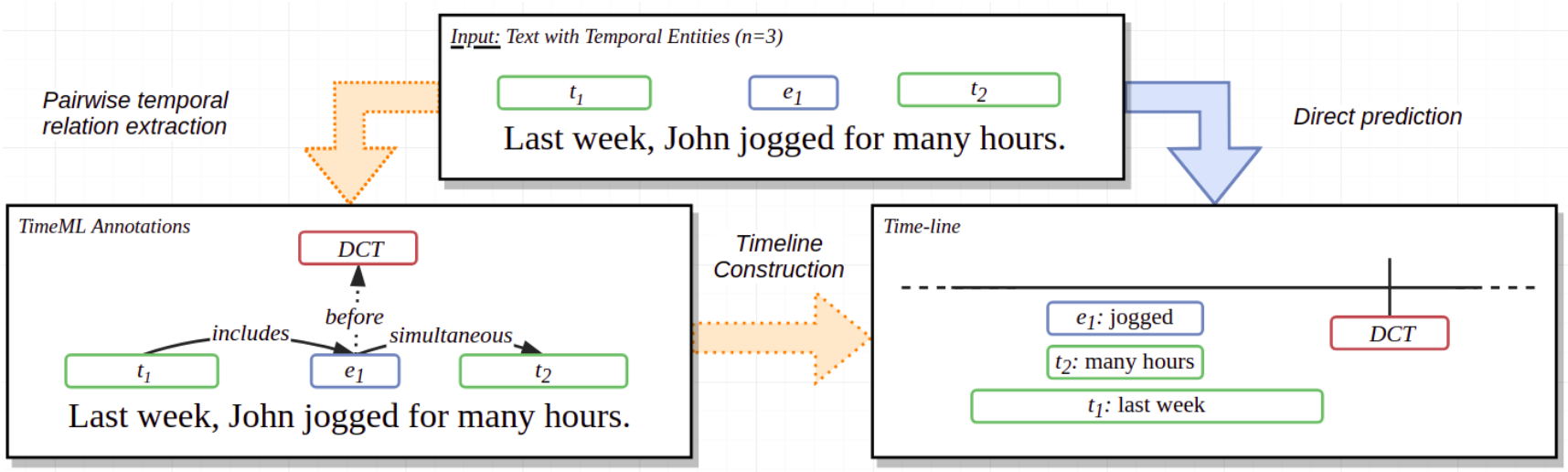


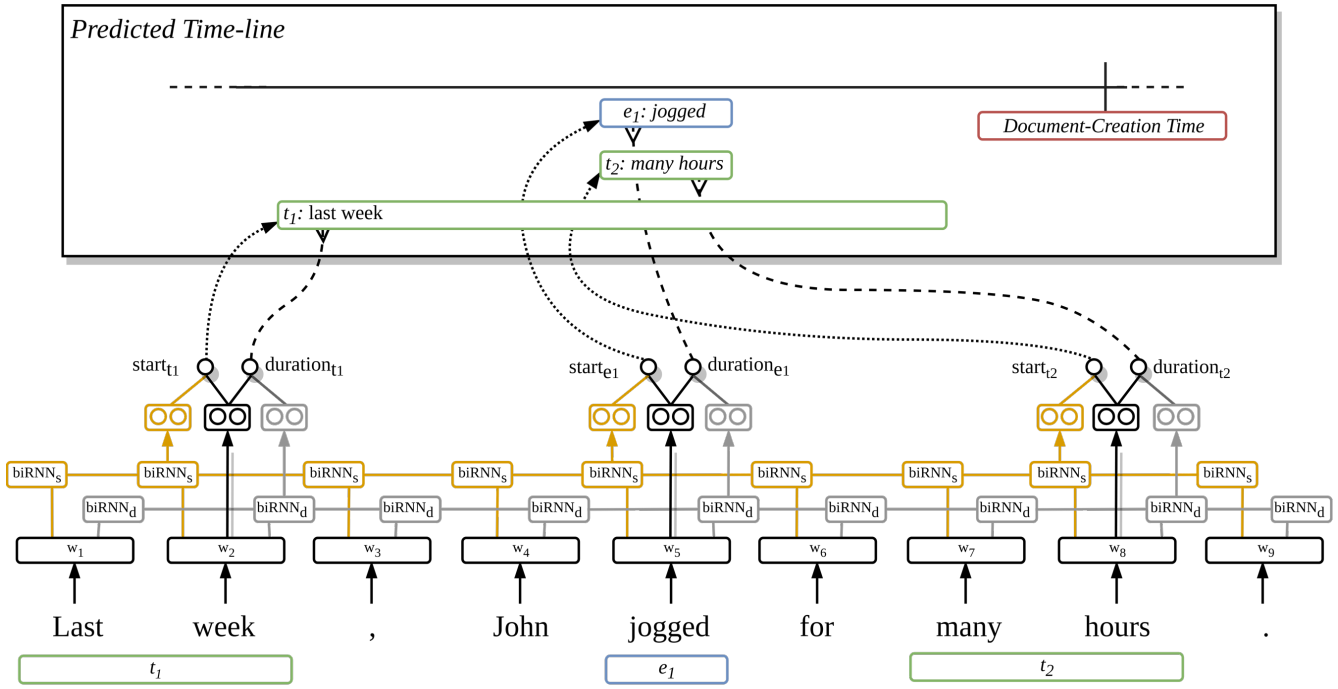
Figure 1: Overview of TIEVis. Clinical events (marked words) are extracted from a clinical report (A) and time periods are estimated for each event. Users can hover over events to bi-directionally highlight them in the report and the visualization (B). An interactive demo is available at: https://augment.cs.kuleuven.be/tievis/

Robin De Croon et al. (2021). TIEVis: A visual analytics dashboard for temporal information extracted from clinical reports. *Companion Proceedings of the 26th International Conference on Intelligent User Interfaces* (pp. 34-36). ACM

Direct prediction of a relative time-line

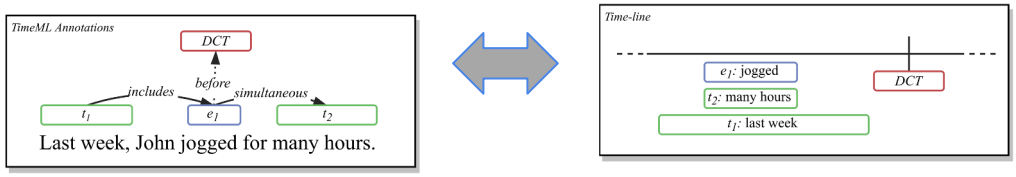


Direct prediction of a relative time-line



Direct prediction of a relative time-line

- Train on data annotated with temporal links
- Use relation between intervals and points to construct a differentiable loss



Allen Algebra	Temporal Links	Point Algebra
X precedes Y Y preceded by X	X before Y Y after X	$e_x < s_y$
X starts Y Y started by X	X begins Y Y begun by X	$s_x = s_y$ $e_x < e_y$
X finishes Y Y finished by X	X ends Y Y ended by X	$e_x = e_y$ $s_y < s_x$
X during Y Y includes X	X is included Y Y includes X	$s_y < s_x$ $e_x < e_y$
X meets Y Y met by X	X immediately before Y Y immediately after X	$e_x = s_y$
X overlaps Y Y overlapped by X	absent[4] absent[4]	$s_x < s_y$ $s_y < e_x$ $e_x < e_y$
X equals Y	X simultaneous Y X identity Y	$s_x = s_y$ $e_x = e_y$

Direct prediction of a relative time-line

Total Time-line Loss $\mathcal{L}_\tau(t,\theta)$:

$$\mathcal{L}_p(\xi|t,\theta) = \begin{cases} \max(\hat{x} + m_\tau - \hat{y}, 0) & \text{if } x < y \\ \max(|\hat{x} - \hat{y}| - m_\tau, 0) & \text{if } x = y \end{cases}$$

- The predicted point $\hat{x}$ should be at least a distance m_τ smaller than predicted point $\hat{y}$ for this loss to be zero
- Predicted point $\hat{x}$ and $\hat{y}$ should be very close to each other at most a distance m_τ away

$$\mathcal{L}_r(r|t,\theta) = \sum_{\xi \in I_{PA}(r)} \mathcal{L}_p(\xi|t,\theta)$$

- Where m_τ = margin, r refers to a TLink, $R(t)$ refers to the set of ground truth T-links of input text t
- Each TLink (relation) loss $\mathcal{L}_r(r|t,\theta)$ is the sum of the point-wise losses $\mathcal{L}_p(\xi|t,\theta)$ of the corresponding algebraic constraints $\xi \in I_{PA}(r)$ from the table in the previous slide
- E.g., combined with negative log likelihood loss of the temporal relation

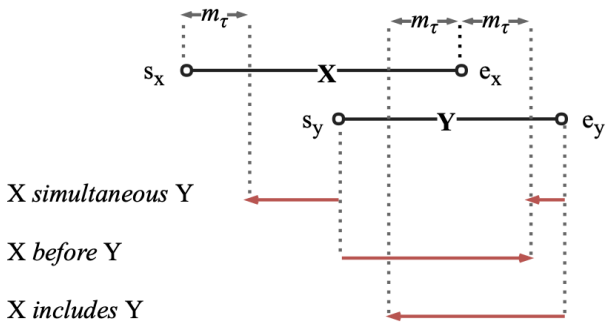


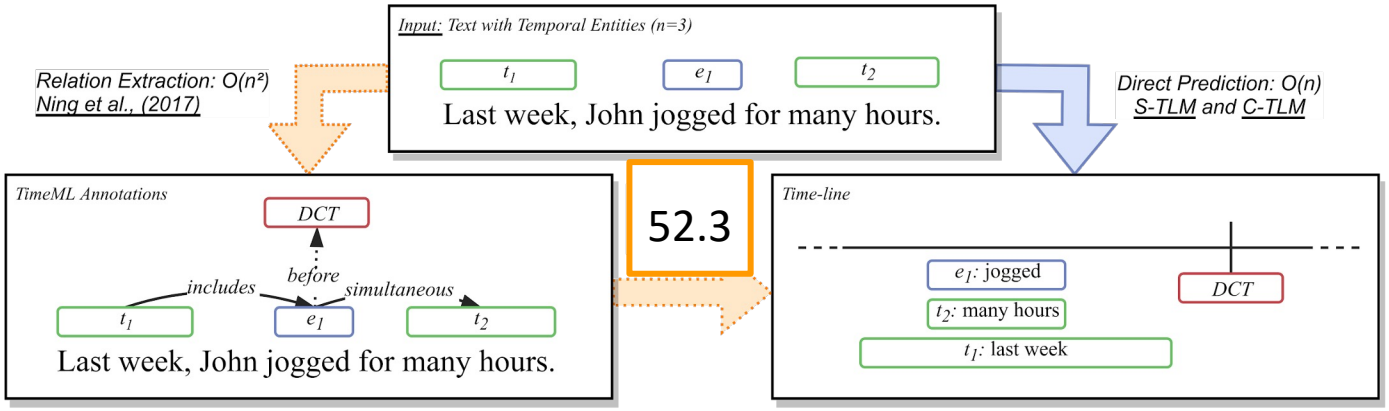
Figure 3: Visualization of the time-line loss L_τ with margin m_τ, for two events X and Y, and TLinks *simultaneous*, *before*, and *includes*. The red arrows' lengths indicate the loss per relation, i.e. how much the points should be shifted to satisfy each relation.

Artuur Leeuwenberg and Marie-Francine Moens (2018). Structured Learning for Temporal Relation Extraction from Clinical Records. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1150–1158. ACL.

Direct prediction of a relative time-line

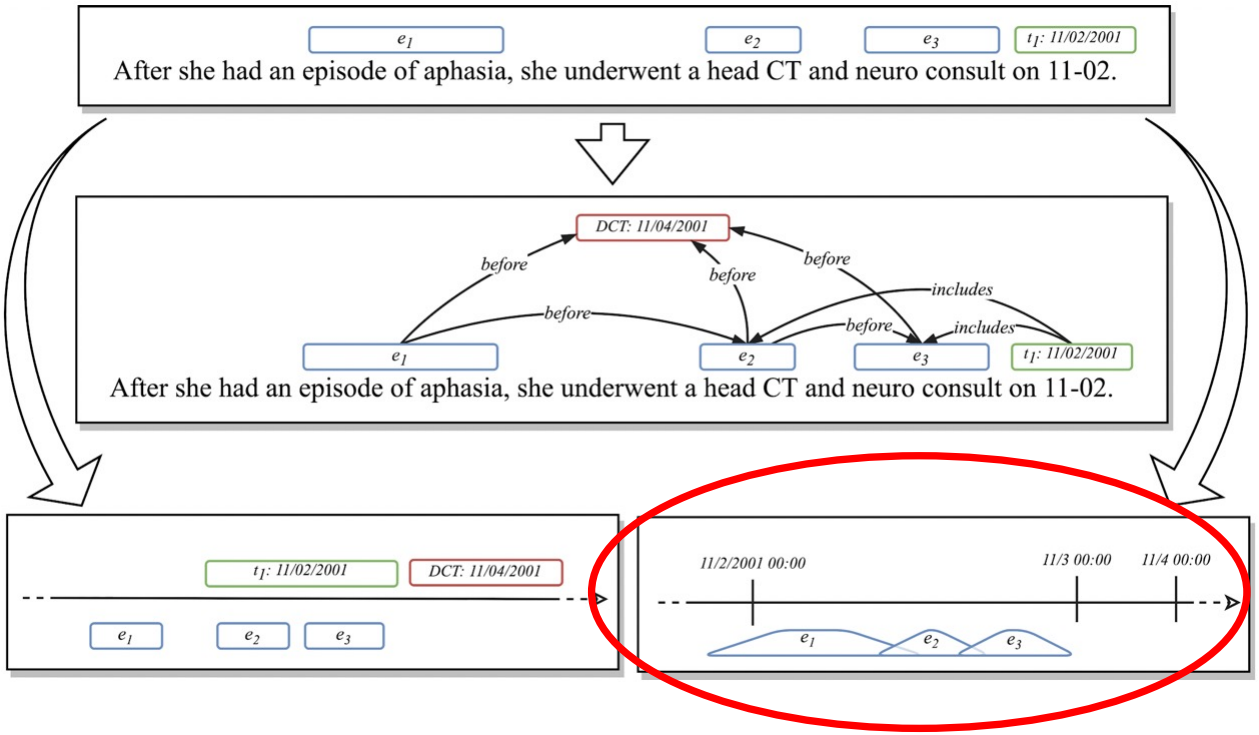
Results: TempEval-3

C-TLM	S-TLM
56.1	50.2



*Temporal awareness F1 measure (UzZaman & Allen, 2011)

Direct prediction of an absolute time-line



$$L_1 = \frac{1}{N}\sum_{i=1}^{N} l(x_i)$$

assume two-piece normal distribution

$$l(x) = \sum_{c}^{C} |\hat{x}_c^{\mu} - x_c^{\mu}| + |\hat{x}_c^{\sigma_l} - x_c^{\sigma_l}| + |\hat{x}_c^{\sigma_r} - x_c^{\sigma_r}|$$

start, duration and end

Artuur Leeuwenberg and Marie-Francine Moens (2022). Towards extracting absolute timelines from English clinical reports. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 28, 2710-2719.