WordsEye - a platform for interactively creating 3D worlds with language

Bob Coyne
coyne@cs.columbia.edu

Historical overview of text-to-graphics

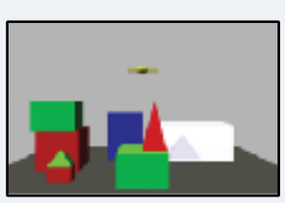
Varieties of Text-to-Graphics Systems

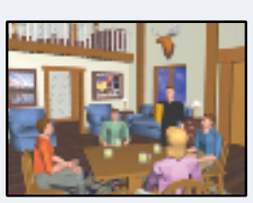
- *Type:* picture/scene construction vs virtual worlds vs animation vs sequential panels
- *Domain-specificity:* e.g. car accident simulations
- Graphical capability and style: simple blocks to 2D images to stylized 2D to photorealistic 3D. Animated characters?
- *Interactivity:* state-based vs stateless? User in the loop? Iteratively refine results? User interface?
- *Language processing and input form:* templates, real-world text, user-specified text.
- *Technique-oriented:* inference, graphic primitives, computational geometry, constraints, machine learning

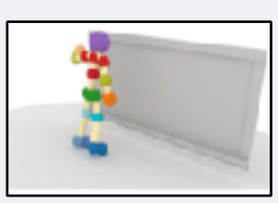

AVDT

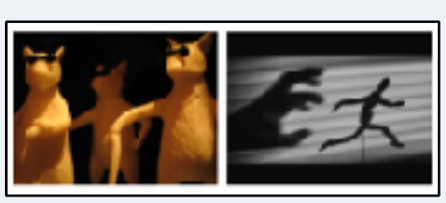

CarSim


Zitnick


Shrdlu


PAR/Jack


PiGraphs


Schwartz

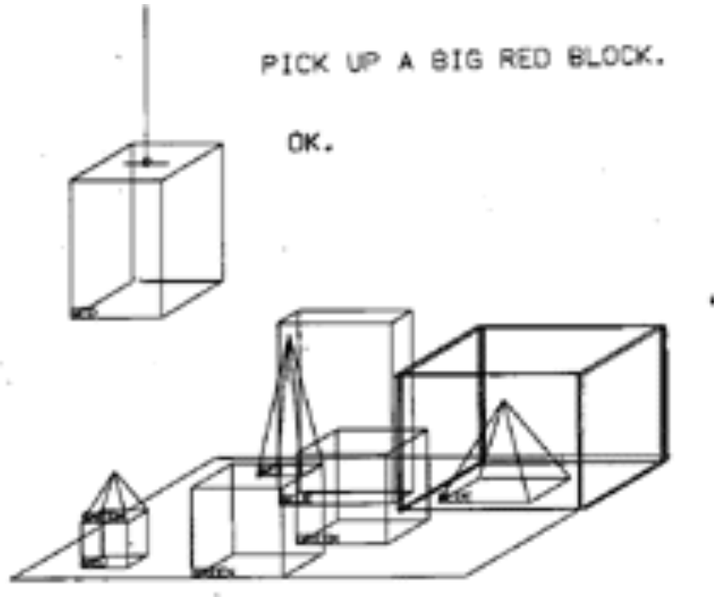

Comic Chat

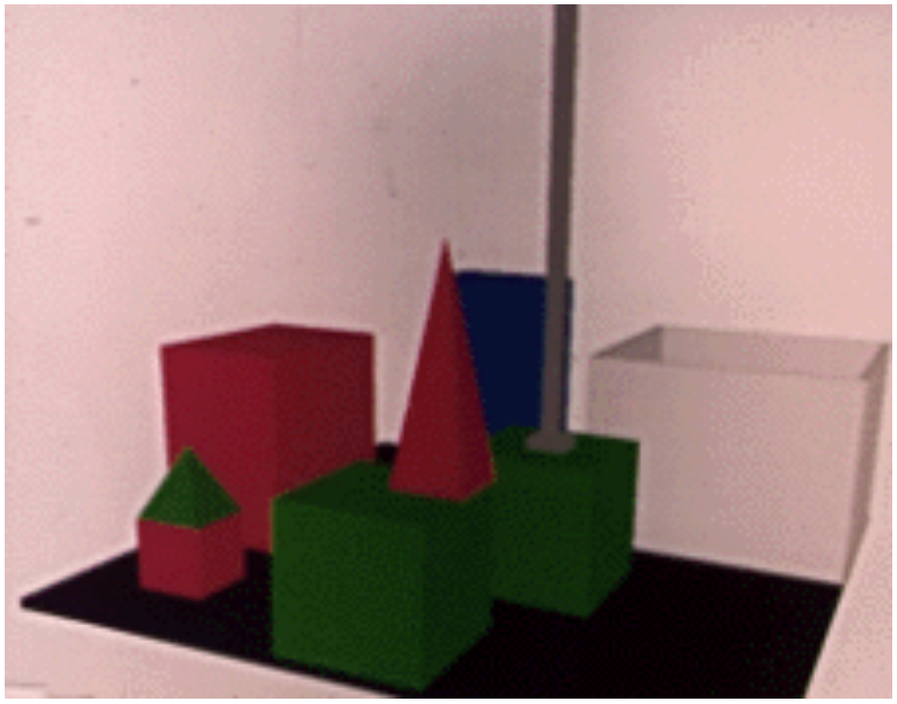
Text-to-Graphics Systems: SHRDLU

- SHRDLU is an early natural-language understanding computer program that was developed by Terry Winograd at MIT in 1968–1970.

- The user carries on a conversation with the computer, moving objects, naming collections and querying the state of a simplified "blocks world", essentially a virtual box filled with different blocks.

- The name SHRDLU was derived from ETAOIN SHRDLU, the arrangement of the letter keys on a Linotype machine, arranged in descending order of usage frequency in English.





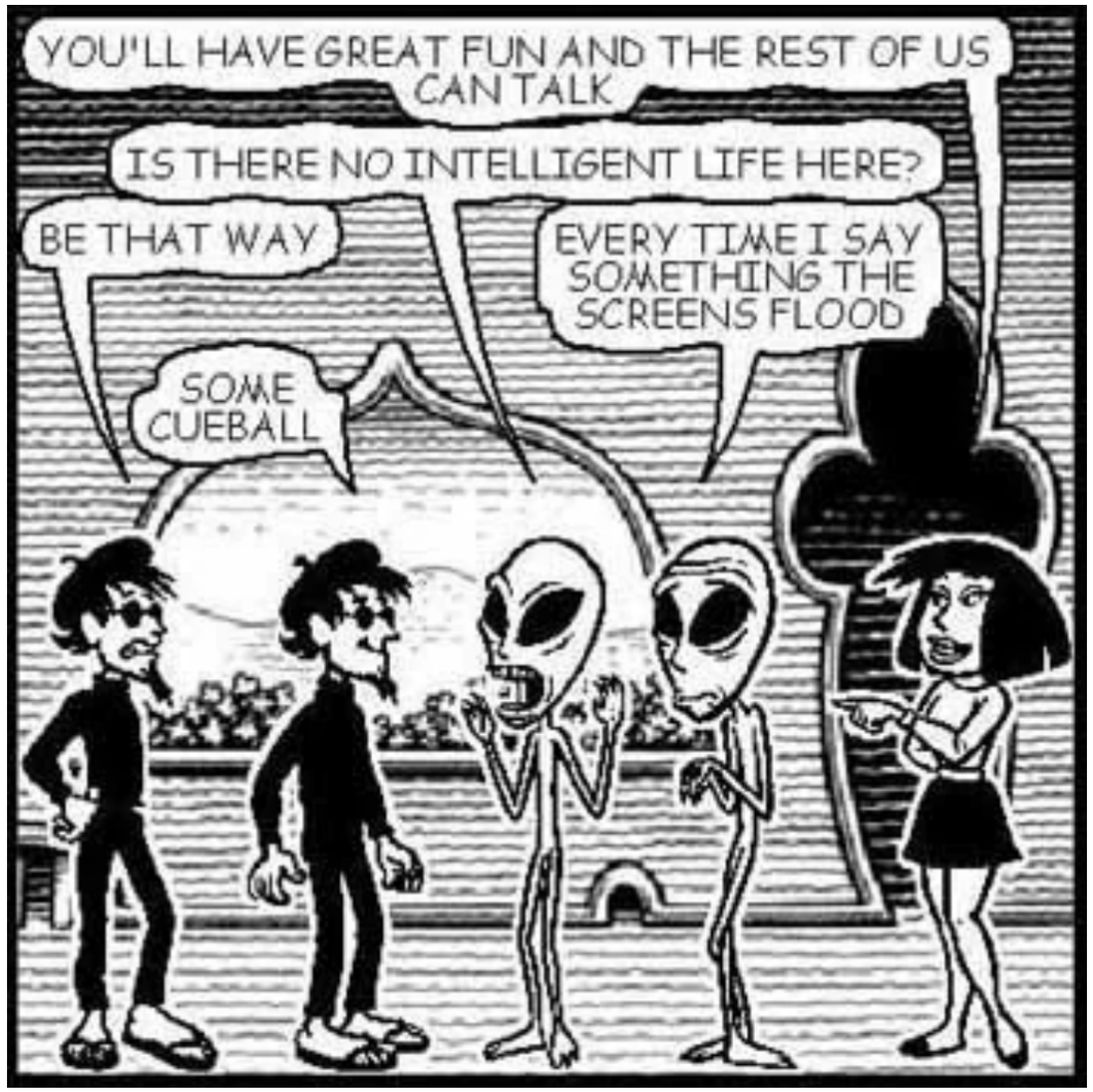
https://en.wikipedia.org/wiki/SHRDLU

Text-to-Graphics Systems: SHRDLU

Person: PICK UP A BIG RED BLOCK.
Computer: OK. (does it)
Person: GRASP THE PYRAMID.
Computer: I DON'T UNDERSTAND WHICH PYRAMID YOU MEAN.
Person: FIND A BLOCK WHICH IS TALLER THAN THE ONE YOU ARE HOLDING AND PUT IT INTO THE BOX.
Computer: BY "IT", I ASSUME YOU MEAN THE BLOCK WHICH IS TALLER THAN THE ONE I AM HOLDING.
Computer: OK. (does it)
Person: WHAT DOES THE BOX CONTAIN?
Computer: THE BLUE PYRAMID AND THE BLUE BLOCK.
Person: WHAT IS THE PYRAMID SUPPORTED BY?
Computer: THE BOX.
Person: HOW MANY BLOCKS ARE NOT IN THE BOX?
Computer: FOUR OF THEM.
Person: IS AT LEAST ONE OF THEM NARROWER THAN THE ONE WHICH I TOLD YOU TO PICK UP?
Computer: YES, THE RED CUBE.
Person: IS IT SUPPORTED?
Computer: YES, BY THE TABLE.
Person: CAN THE TABLE PICK UP BLOCKS?
Computer: NO.
Person: CAN A PYRAMID BE SUPPORTED BY A BLOCK?
Computer: YES.

Text-to-Graphics Systems: Comic Chat

- Comic Chat [Kurlander *et al*., 1996] converts chat streams to comic book panels with the participants' dialog assigned to different 2D graphical comic characters.
- The emphasis of this system is on the presentational style and graphical layout of the resulting cartoon panels. The characters' gestures and facial expressions are determined by a set of simple rules.
- For example, greeting words in the chat will cause the chat character to be put into a waving pose. And self-references (such as *I* or *I'll*) will cause the character to point to itself.



https://kurlander.net/DJ/Projects/ComicChat/resources.html

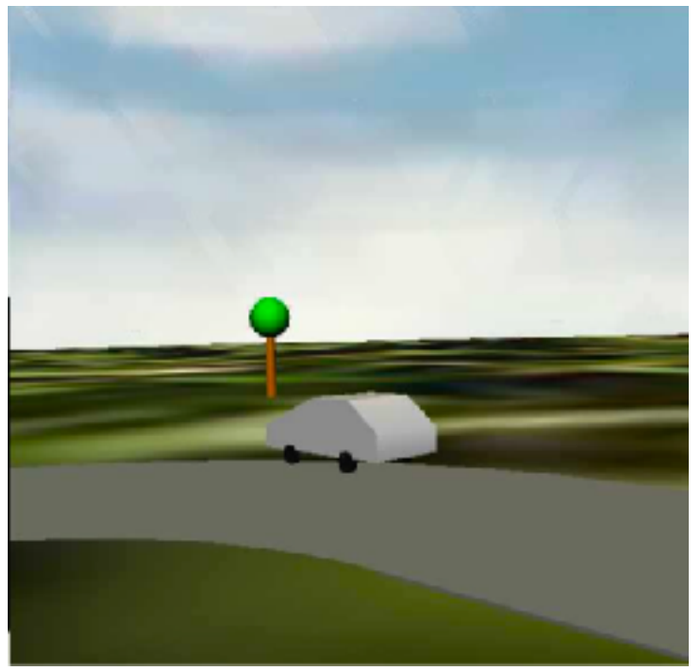
Text-to-Graphics Systems: Carsim

Johansson et al.: Carsim (2004): A system to visualize written road accident reports as animated 3d scenes

Corpus of 200 accident reports from Swedish newspapers.

- Length varies from couple sentences to over a page
- Style varies from very detailed to implicit
- Additional reports from STRADA database

Example: *A fatal accident took place tonight south of Vissefjarda on Road 28. A car carrying two persons departed from the road in a left-hand curve and crashed at a high speed into a spruce. The passenger, who was born in 1984, died. The driver, who was 21 years old, is severely injured and is taken care of in a hospital. The police suspects that the car they were traveling in, a new Saab, was stolen in Emmaboda and will investigate it today.* (translated)



Text-to-Graphics Systems: Zitnick

[Zitnick *et al*., 2013] learns and creates scenes from 2D clip art. Trained and tested using a dataset of 10,000 images of children playing outside.

Trained on 60,000 sentences that described the scenes in different ways. Amazon Mechanical Turkers create scenes by positioning the clip art. 80 pieces of clip art representing 58 different objects for common objects such as people, plants, and animals.

For text input, each sentence is represented by a tuple, containing a primary object, a relation and optionally a secondary object.

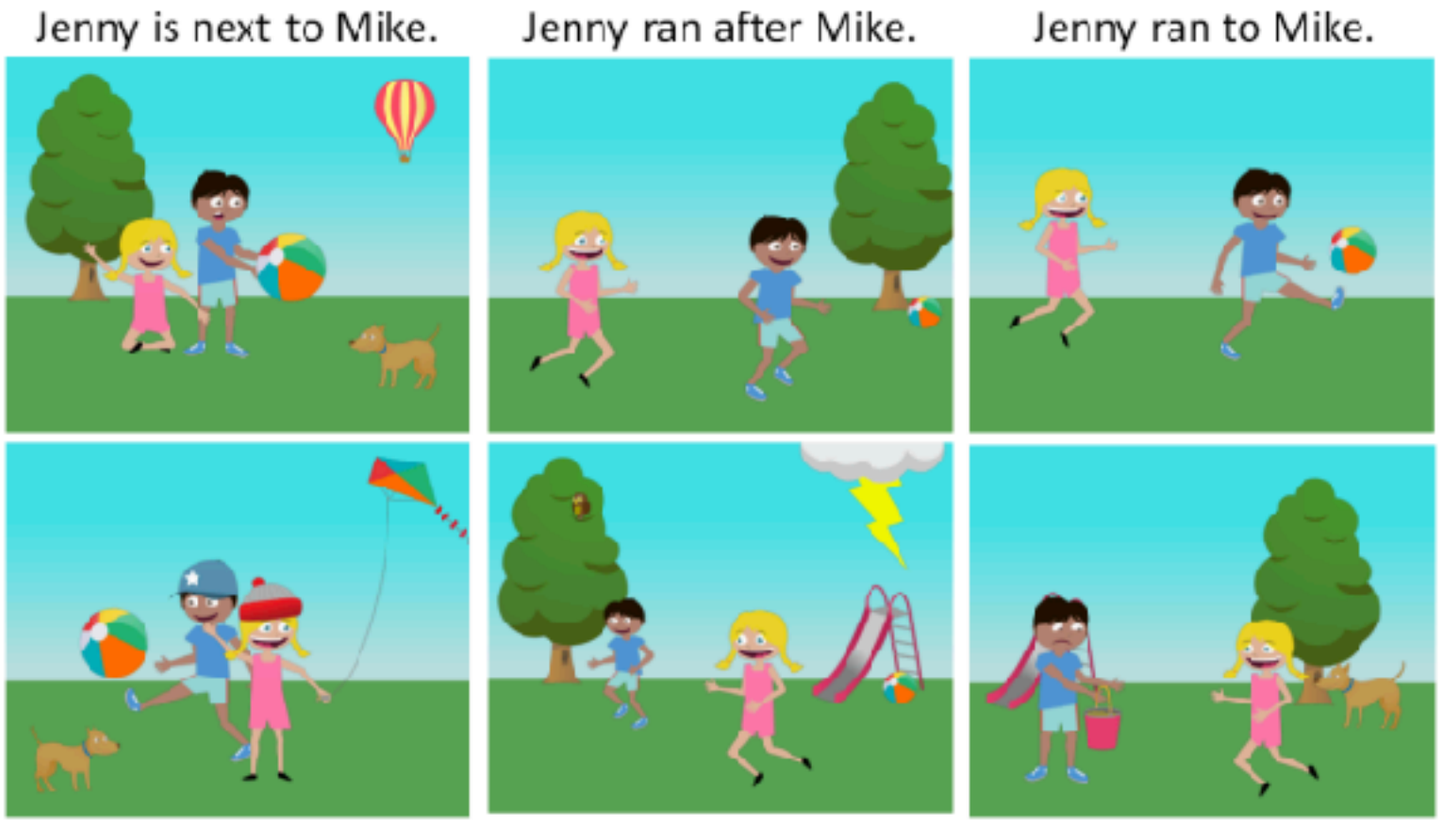
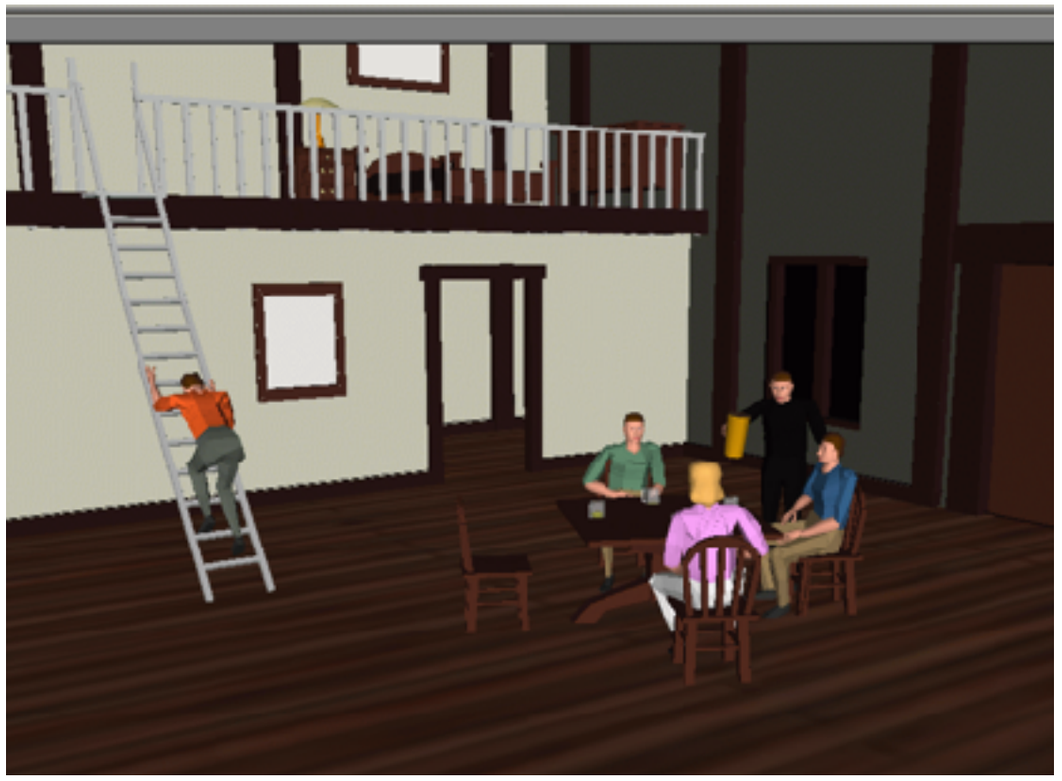


Figure 1: Three example sentences and scenes. Notice how subtle changes in the wording of the sentences leads to different visual interpretations.

http://larryzitnick.org/publication/ZitnickSent2SceneICCV13.pdf

Text-to-Graphics Systems: PAR

- The PAR System is a graphical semantic framework that uses language to control animated characters in a closed pre-constructed virtual environment.
- The PAR (Parameterized Action Representation) framework represents AGENT, SOURCE, GOAL, PATH, and other verb arguments. A PAR gives a complete representation of the action. Each action also has applicability conditions, pre-conditions, and post-assertions.
- Input is parsed and used to instantiate a PAR. References are grounded in objects in the environment. The system supports a limited set of actions (walking, sitting down on a chair or bed, standing up, talking to others, climbing a ladder, opening a door, shaking hands, drinking).



https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=f6ed2e59d7dbd451e12e609344a1f0a35879eeb9

Current approach - 2D diffusion models

- Trained on labeled images (not 3D — works on pixels)
- Produces visually compelling and relevant output for almost any input.
- Very successful both technically and commercially. Midjourney has achieved over $200 million in revenue with 40 employees, without any external investors.
- Intellectual property issues and lawsuits
- No explicit objects or world model (making it difficult to control details and spatial relations).
- Platforms: Midjourney, Dall-E, Stable Diffusion, and others

Midjourney

A Midjourney-created image of Pope Francis wearing a puffer jacket, which went viral in 2023

Prompt: A cowboy wearing a tuxedo on the moon

Stable Diffusion

Stable Diffusion is open source

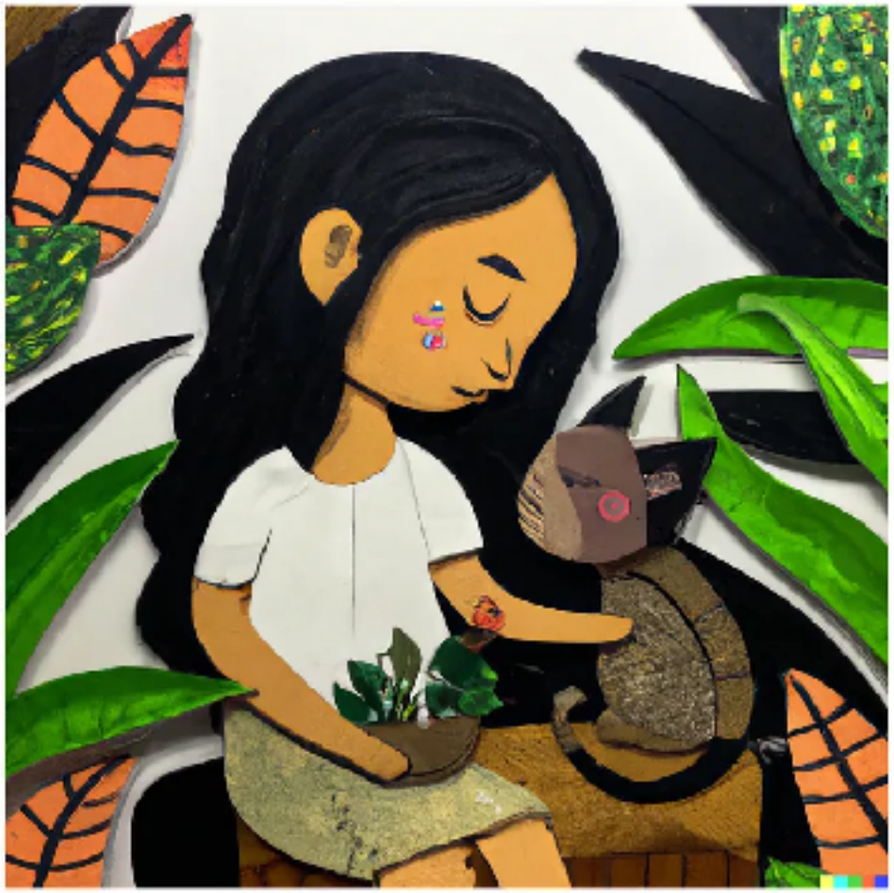
Prompt: *"a photograph of an astronaut riding a horse"*

https://en.wikipedia.org/wiki/Stable_Diffusion

Dall-E

Dall-E 3

Dall-E 2



Prompt: *A paper craft art depicting a girl giving her cat a gentle hug. Both sit amidst potted plants, with the cat purring contentedly while the girl smiles. The scene is adorned with handcrafted paper flowers and leaves.*

Sora - AI generated video

Prompt: Beautiful, snowy Tokyo city is bustling. The camera moves through the bustling city street, following several people enjoying the beautiful snowy weather and shopping at nearby stalls. Gorgeous sakura petals are flying through the wind along with snowflakes.

https://openai.com/sora

Sora - AI generated video

Prompt: Five gray wolf pups frolicking and chasing each other around a remote gravel road, surrounded by grass. The pups run and leap, chasing each other, and nipping at each other, playing.

https://openai.com/sora

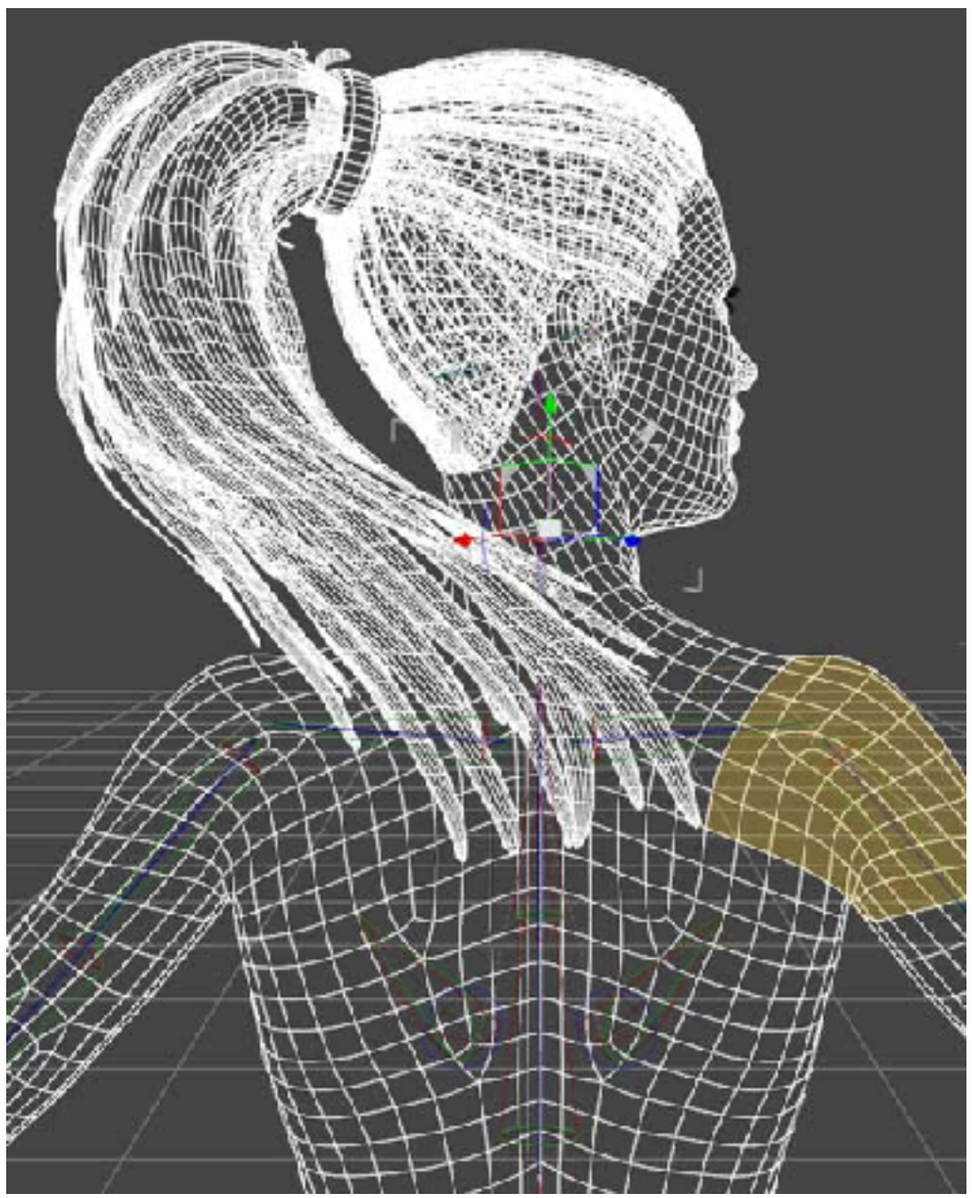
Diffusion models — IP issues

- Generative AI images not protected by copyright law
  https://www.reuters.com/legal/ai-created-images-lose-us-copyrights-test-new-technology-2023-02-22/
- Generative AI images can plagiarize existing art. For example, Midjourney produced these recognizable images of The Simpsons.

https://spectrum.ieee.org/midjourney-copyright

Let’s talk about the benefits of 3D…

3D graphics



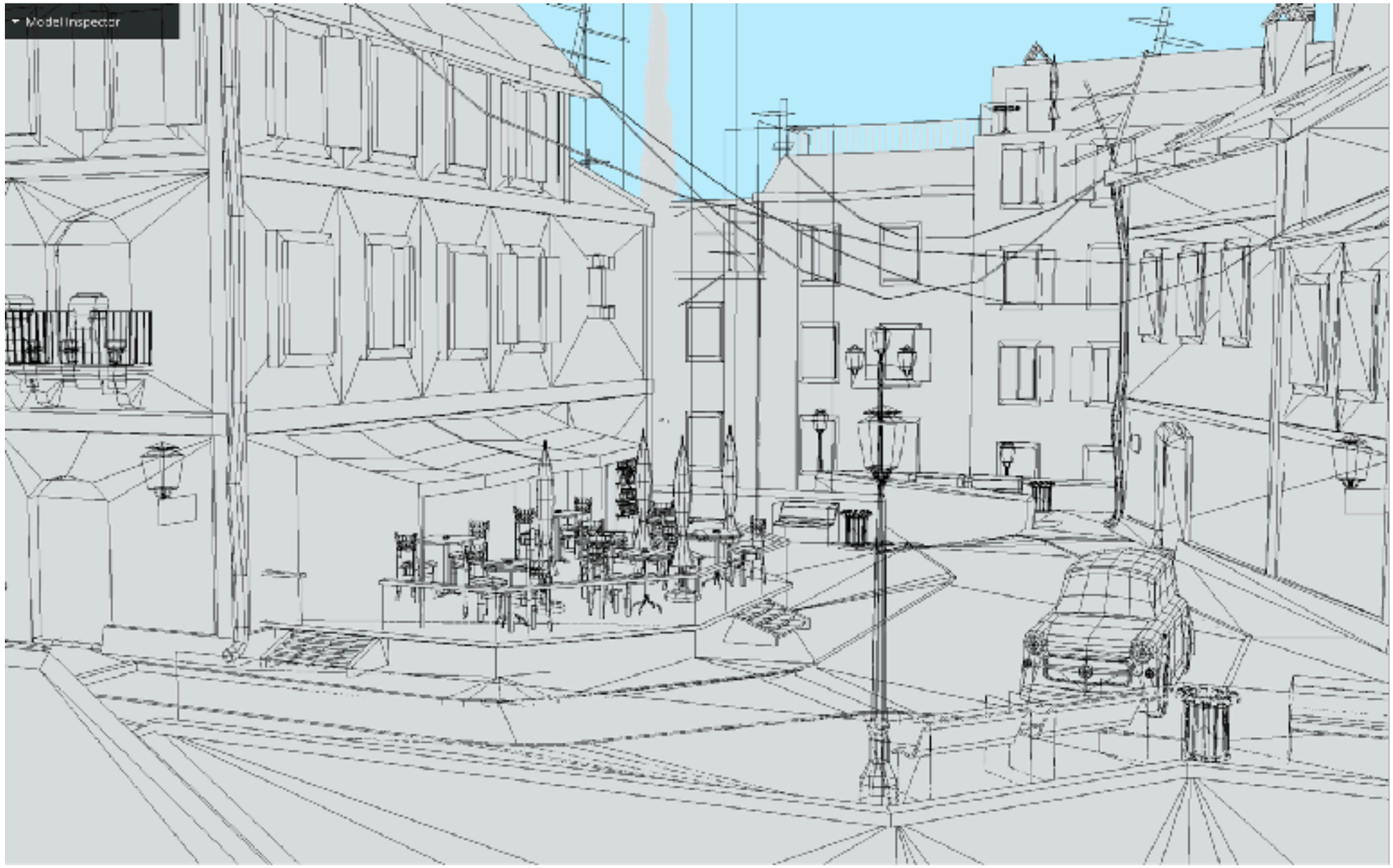
Wire-frame 3D polygonal mesh

Objects represented by 3D meshes.
Used in computer games, films, special effects, product design, architecture, etc.

Rendered 3D model

3D graphics

Polygonal meshes to represent 3D objects



3D graphics

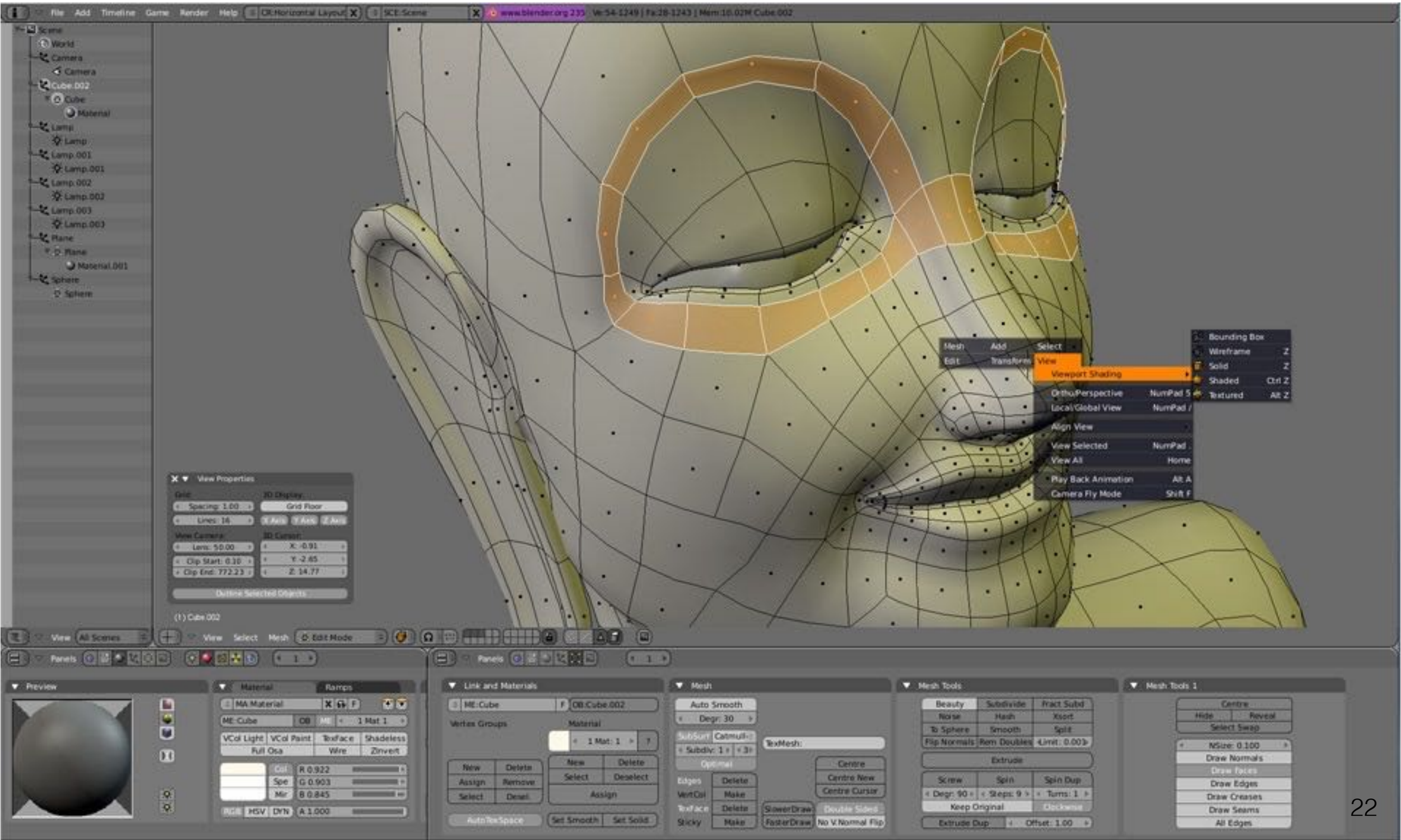
Rendered scene generated from polygon mesh, surface attributes, and light sources

3D graphics

It's an explicit world model — can be re-rendered (or animated) from ANY point of view.
Objects can be arbitrarily and independently modified and combined

3D graphics can depict anything imaginable, but ….

3D tools are complex



WordsEye — Using language to create interactive 3D worlds

WordsEye on the web

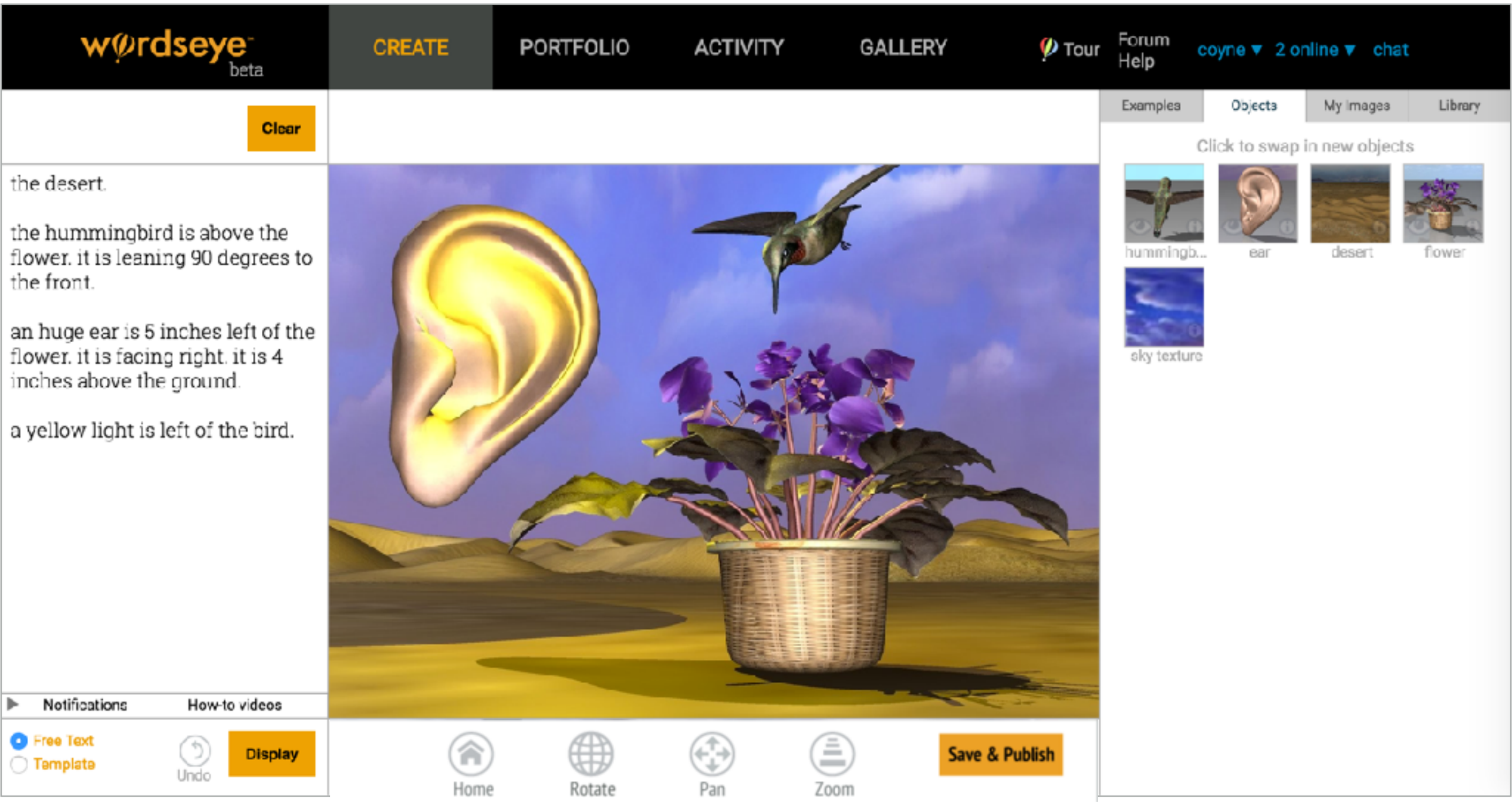
WordsEye on the web (2015-2021)

- 80K registered users
- 28K rendered scenes posted to Gallery
- 500 users with > 10 sessions
- A few hard-core users online 100+ hrs/month

Problem: While some users became were able to create compelling scenes, others had difficulties.

- Non-real-time camera made it hard to adjust viewpoint
- Exclusively textual specification of spatial relations was limiting

WordsEye web interface



WordsEye user examples

Sunset on the Marsh

Input text: *a pond is in the swamp. sky. a dark wood boat is 10 feet behind and -12 feet left of the pond. it faces left. a person is sitting in the boat. the person faces right. the boat's oar is black. 2 lights are 4 feet right of and above the person. camera light is black. sun is linen.*

@nheiges

Wolf moon

Input text: *sky is black.ground is invisible. a 300 inch tall moon.a 250 inch tall black wolf is in front of the moon.the wolf is facing west.*

@kawe

WordsEye user examples

Fishing

Input text: *a carving. fish is -2 inch above carving. fish is facing southeast. fish is leaning 85 degrees to right. mauve 2.6 foot tall compass is -1 inch right of carving. compass is facing right. 3.2 foot tall bordeaux wine mauve sun symbol is -3 inch in front of carving. ground is shiny pond green. 6 foot tall white first mannequin is 3.5 foot behind carving. mannequin is facing up. mannequin is facing right. small arrow is 2 inch in fish. arrow is leaning 90 degrees to front. 6 foot tall clear second mannequin is 1.3 foot above first mannequin and -5 foot to left….*

@tane69

"Excuse me Miss, I'm looking for the Brunt Ice Shelf..."

Input text: *Statue of Liberty is -33 feet above a shiny lake. The lake is 50 feet wide water. New York backdrop. Camera light is black. A huge duck is -17.1 feet above and -2 feet right of and behind the Statue of Liberty. It is facing southwest.*

@hedgehog1965

WordsEye user examples

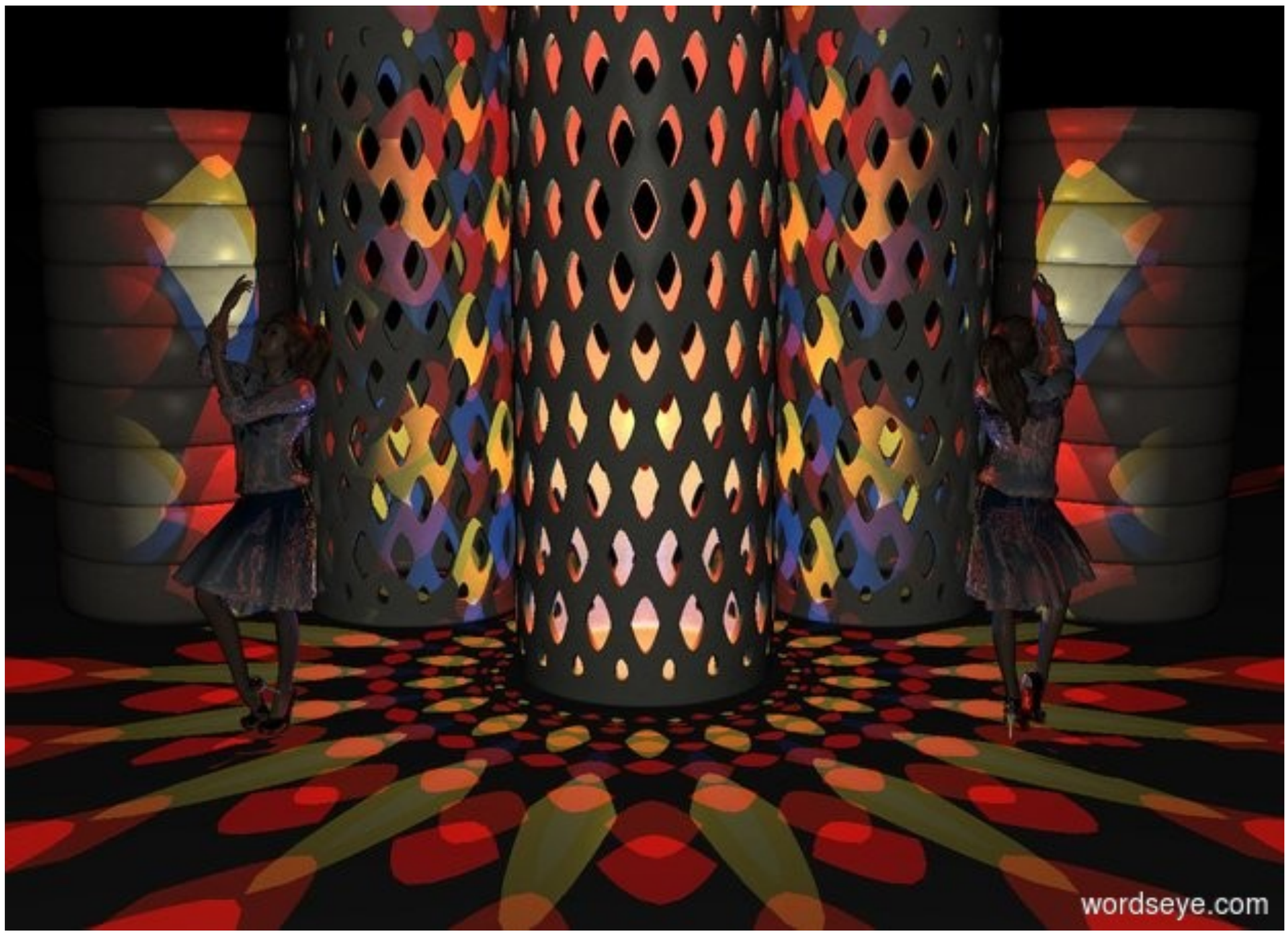
Waiting for the 13:15

Input text: *a elephant. A man is 1 feet right of the elephant. A train is in front of the man.it is facing left. The train is right of the man. shiny ground.*

@watcher570

WordsEye user examples

Disco Lights



Input text: *a 1st vase. it is night. camera light is 30% dim. a tiny lemon light is -1.75 feet above the vase. a tiny fire orange light is -.75 feet above the vase. a tiny swimming pool blue light is -2.3 feet above the vase. a tiny red light is -.25 feet above the vase. a 2nd vase is behind and right of the vase. a 3rd vase is right of the vase. a 4th vase is left of and behind the 1st vase. a 5th vase is left of the vase. a 1st tiny shiny person is 1 feet in front of and right of the vase. a 2nd tiny shiny person is 1 foot in front of and left of the 3rd vase. she faces back.*

WordsEye user examples

Walk Into My Parlor

Input text: *a mike. stage backdrop. a huge fly is -.85 feet above and -3 feet right of and -.1 feet in front of the mike. it leans 60 degrees to the northeast. the wing of the fly is 1 inch tall shiny [texture].the body of the fly is scales. it faces northeast. a giant metal web is in front of and -3 feet above the mike. it leans back. a 3.7 feet tall and 3 feet wide wood table is -1.8 feet right of and -7 feet above the mike . it leans 10 degrees to the southeast. it faces southwest. a very huge shiny spider is -.37 feet above the table. a silver plate is .11 feet left of and .11 feet in front of and -.43 feet above the spider. it leans 7 degrees to the southeast.*

@nanook

WordsEye user examples

Everyone Needs a Lucky Pudding This Christmas

Input text: *A large shiny shilling is in a cake. It is leaning 10 degrees to the back. Fireplace backdrop. A gingerbread man is 2 inch left of the cake. It is facing east. Backdrop is 20% shiny. Sky is Christmas. Sky is leaning front. The cake is on a Christmas table. A woman is -9 inch right of and -4.6 feet above and -1 foot behind the cake. She is facing west. She is leaning front. Camera light is black. An orange light is above and behind and left of the cake. A lemon light is in front of and above the shilling. A 9 inch high glass is behind and -3.7 inch right of the cake.*

@hedgehog1965

WordsEye user examples

You're so vain

@nanook

Input text: *a entryway. a man is -16 feet left of and -9 feet in front of and -8.78 feet above the entryway. a 9 feet tall and 12 feet long marble wall is left of the entryway. it faces left. it is noon. sun is dim ochre brown. ambient light is delft blue. a fjord blue light is right of and in front of the man. 1st woman is 6 feet left of and 1 feet behind the man. she faces the man. the dress of the woman is 6 inch tall [texture]. a 6 feet wide yacht painting is -6 feet above and -7 feet behind and 1 inch right of the wall. it faces right. the frame of the painting is texture. a very tiny apricot ghost is -1.38 feet above and -9.3 inch in front of and -1.2 feet left of the man. it leans back. a wine bottle is -3.9 feet above and -8 inch left of and -1 feet in front of the man. it leans 45 degrees to the left. it faces southwest. 2nd woman is 1 feet in front of the 1st woman. she faces the 1st woman. a storm blue light is left of the 1st woman. a .49 feet tall and 1.2 feet deep and .8 feet wide pewter gray fedora is -.52 feet above and -1.44 feet behind and -1.6 feet left of the man. it faces southeast. it leans to the front.*

WordsEye World

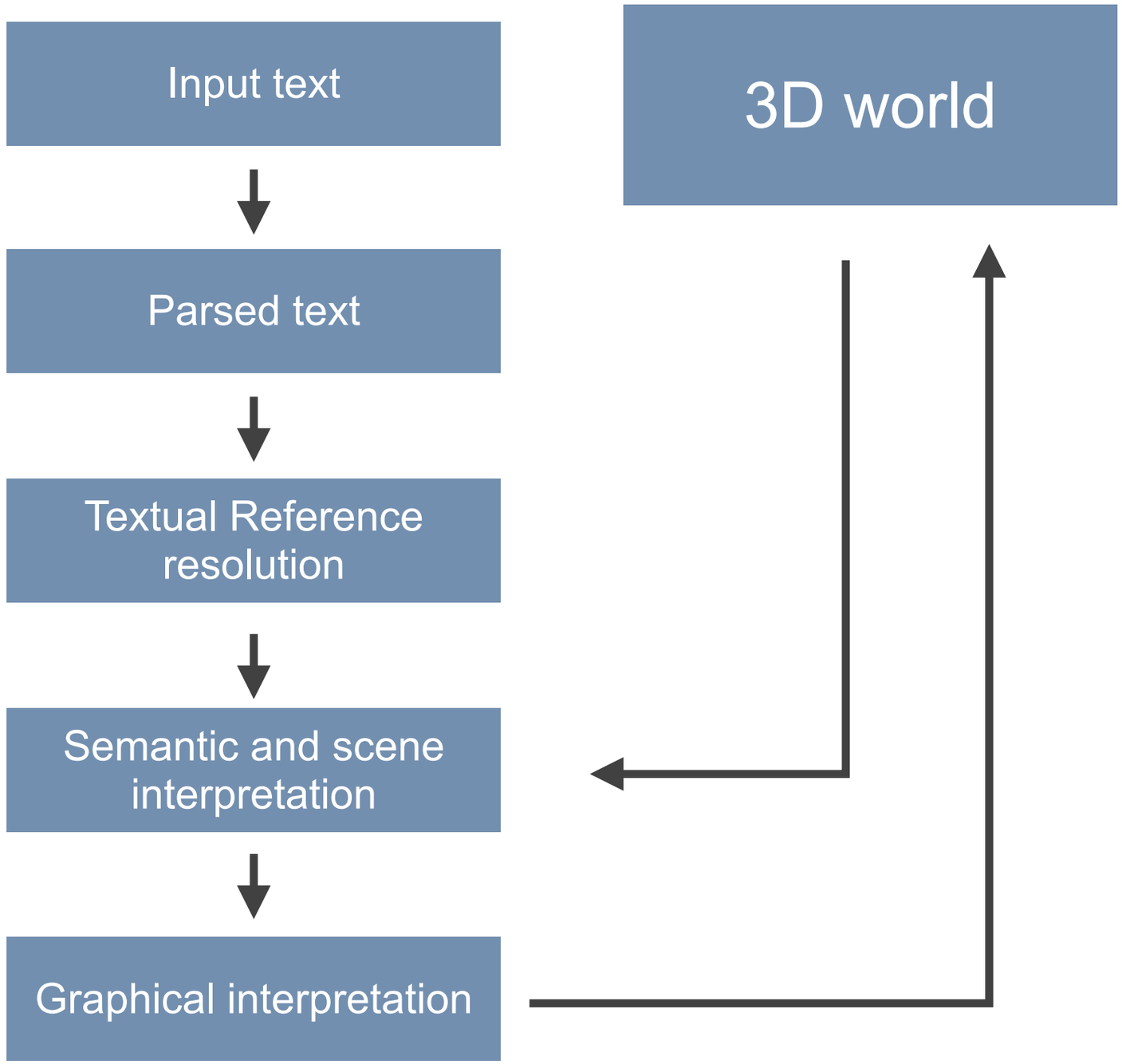
- Work in-progress
- Based on Unity game engine for real-time 3D on the web
- Create immersive and interactive 3D worlds and experiences vs rendered picture as output
- Iterative scene building and workflow — Refer to existing objects in scene vs deriving scene solely from a single block of text.
- Include the ability to point at locations (“here” and “there”) to integrate with textual specifications.
- Specify simple real-time actions in addition to the scenes themselves

WordsEye World demo

https://www.wordseyeworld.com/

How it works

WordsEye pipeline

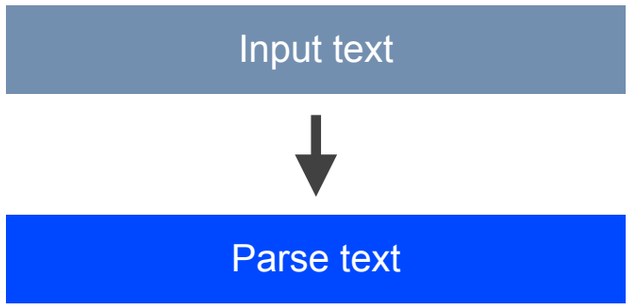


WordsEye pipeline

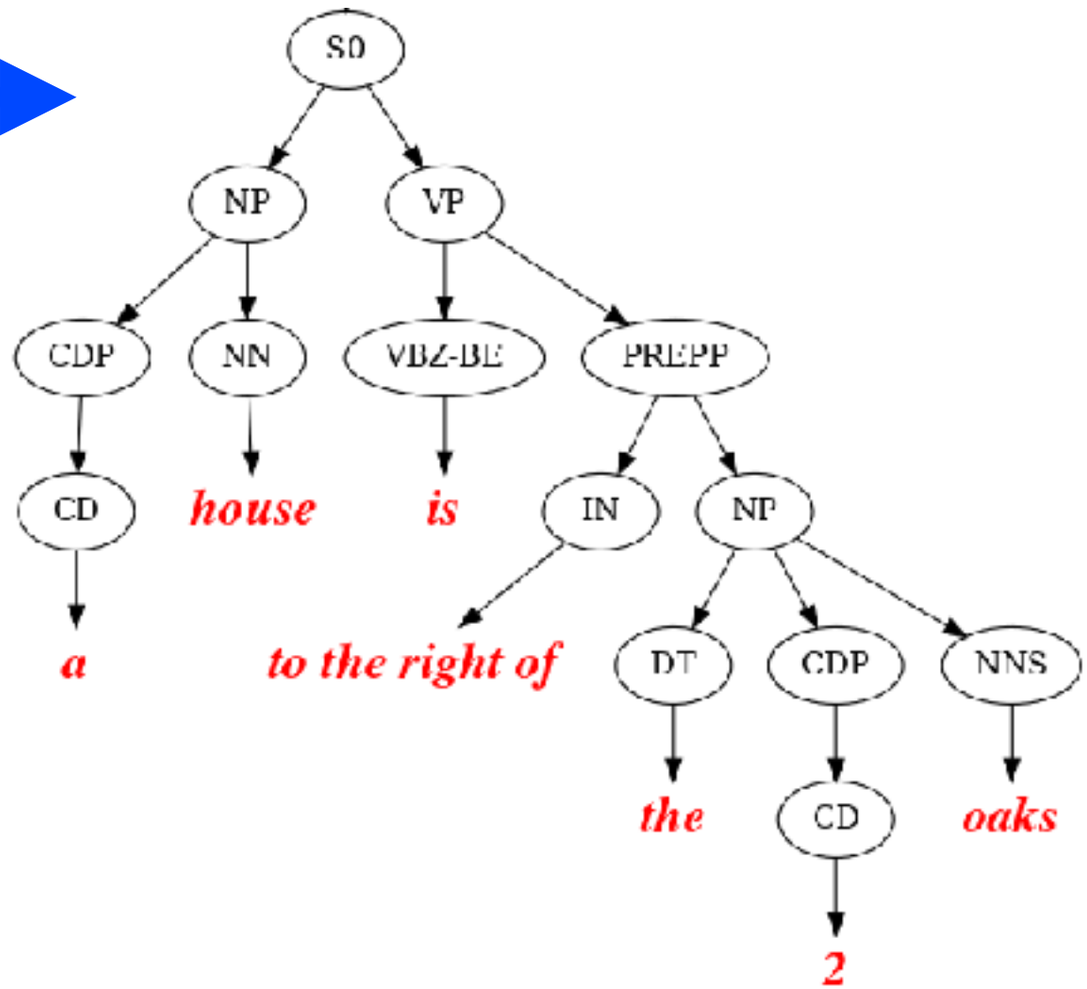
Input text

“A house is to the right of the two oaks.
There is a blue horse 3 feet in front of it.”

WordsEye pipeline



"A house is to the right of the two oaks.
There is a blue horse 3 feet in front of it."



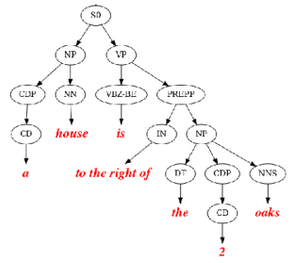
WordsEye pipeline

Input text

Parsed text

Textual Reference resolution

*"A **house** is to the right of the two oaks.*
*There is a blue horse 3 feet in front of **it**."*



Merge references between ***it*** and ***house***

WordsEye pipeline

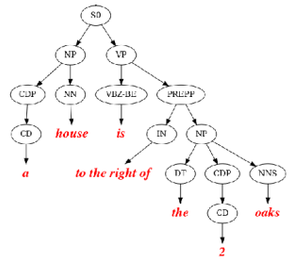
Input text

Parsed text

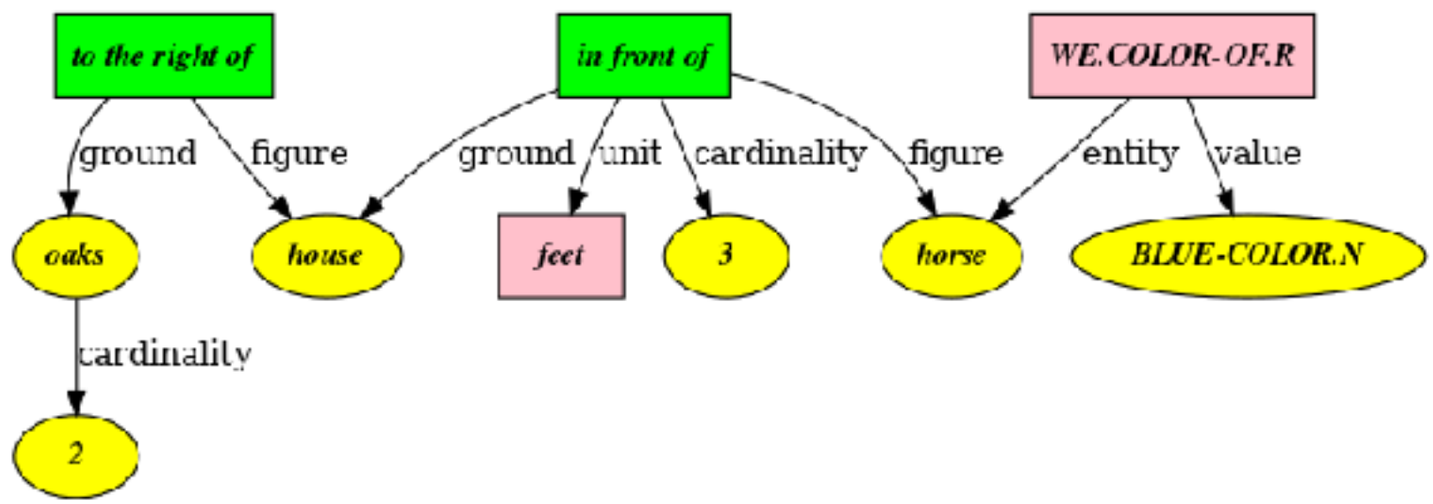
Textual Reference resolution

Semantic and scene interpretation

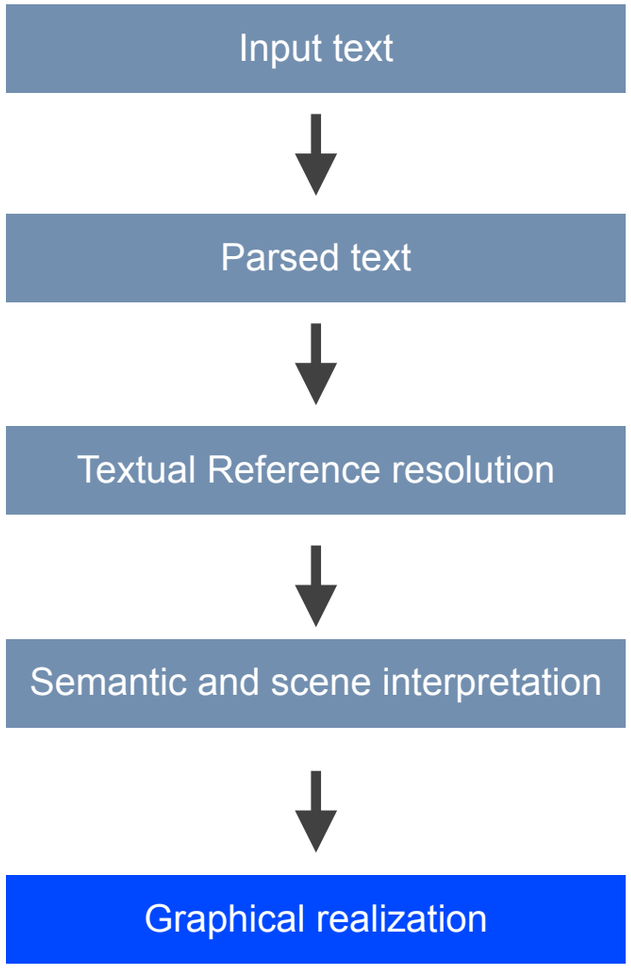
*"A house is to the right of the two oaks.
There is a blue horse 3 feet in front of it."*



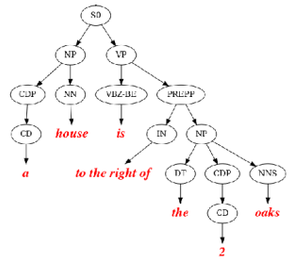
Merge references ***it*** and ***house***



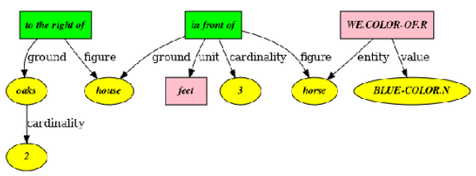
WordsEye pipeline



"A house is to the right of the two oaks.
There is a blue horse 3 feet in front of it."



Merge references ***it*** and ***house***



Modify scene (new input text)

*“An elephant is in front of the **first tree”***

Modify scene (new input text)

*“An elephant is in front of the **first tree”***

↓

Parse & interpret new text

Modify scene (new input text)

*"An elephant is in front of the **first tree"***

↓

Parse & interpret new text

↓

Reference resolution

New: *Elephant*

Merge: **"***first tree"* with scene's oak tree

Modify scene (new input text)

*"An elephant is in front of the **first tree"***

Parse & interpret new text

Reference resolution

New: *Elephant*

Merge: **"***first tree"* with scene's oak tree

Update semantics/graphics

Semantic representation

Meaning represented as **concepts** and **semantic relations** applied to them.

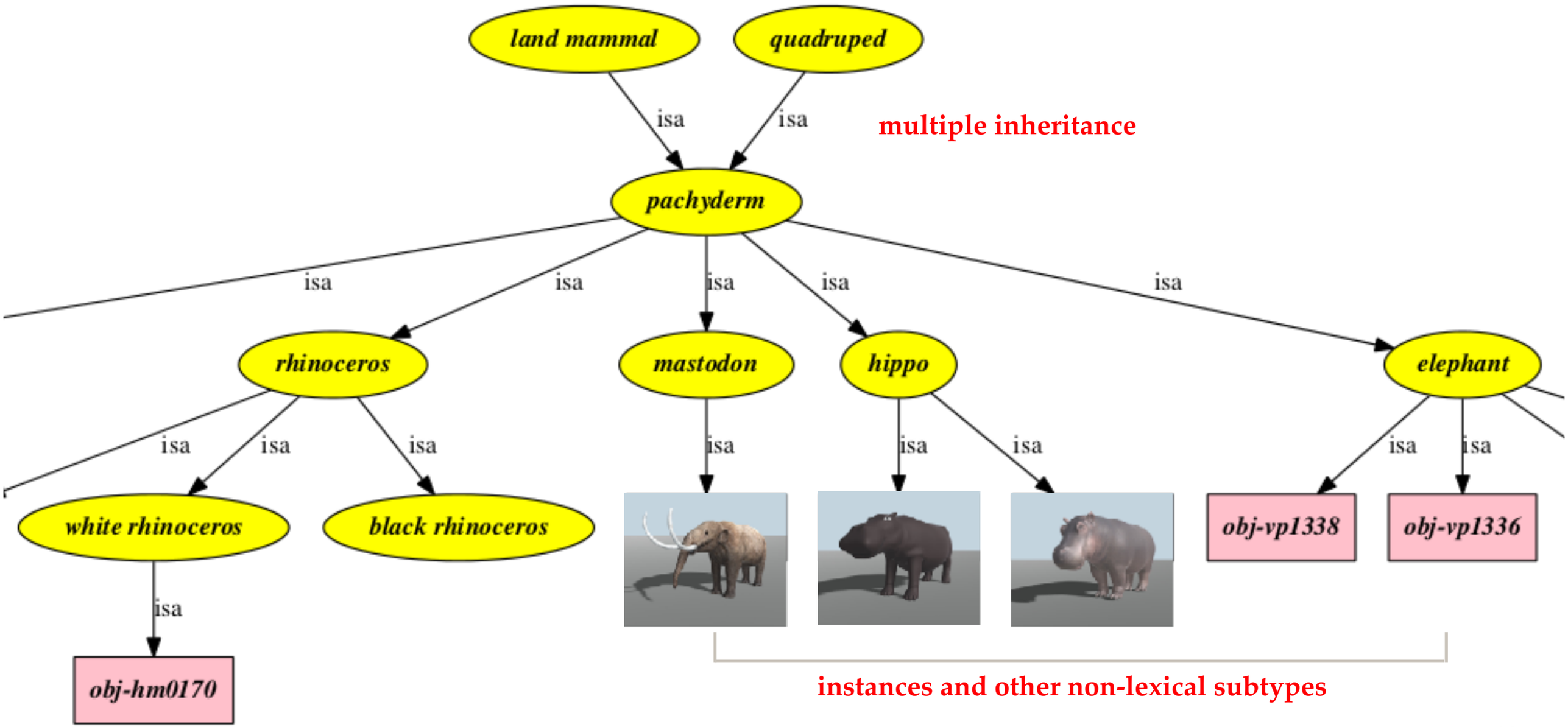
Concepts — represent nouns or other referable entities

- Objects, collections, types, events, anonymous instances …
- Structured in an IS-A hierarchy.

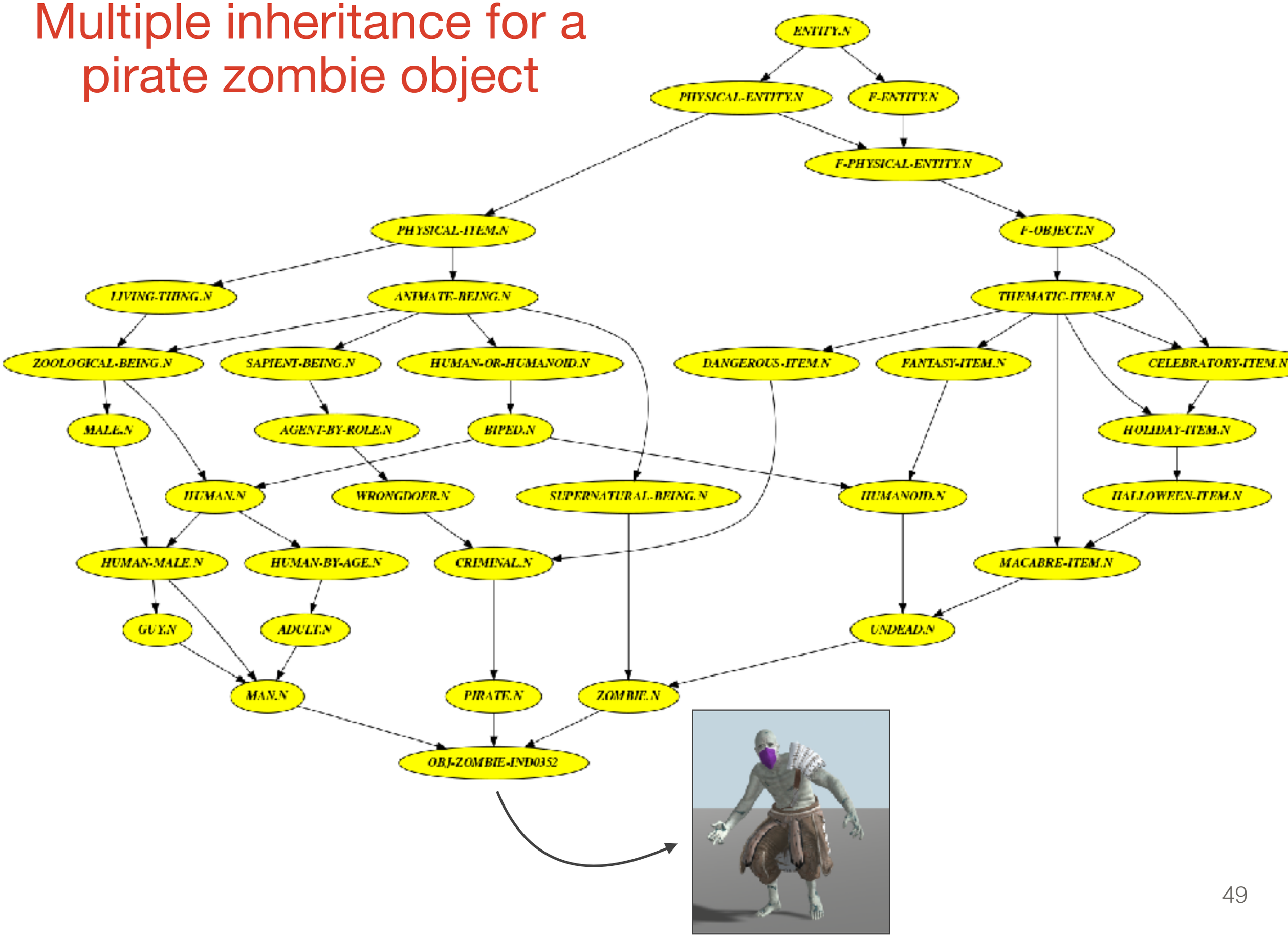
Entities — instantiated concepts representing individual objects

Semantic Relations — applied to concepts and entities. They represent verbs, adjectives, prepositions, adverbs, and graphical/spatial relations.

Concepts and lexical items

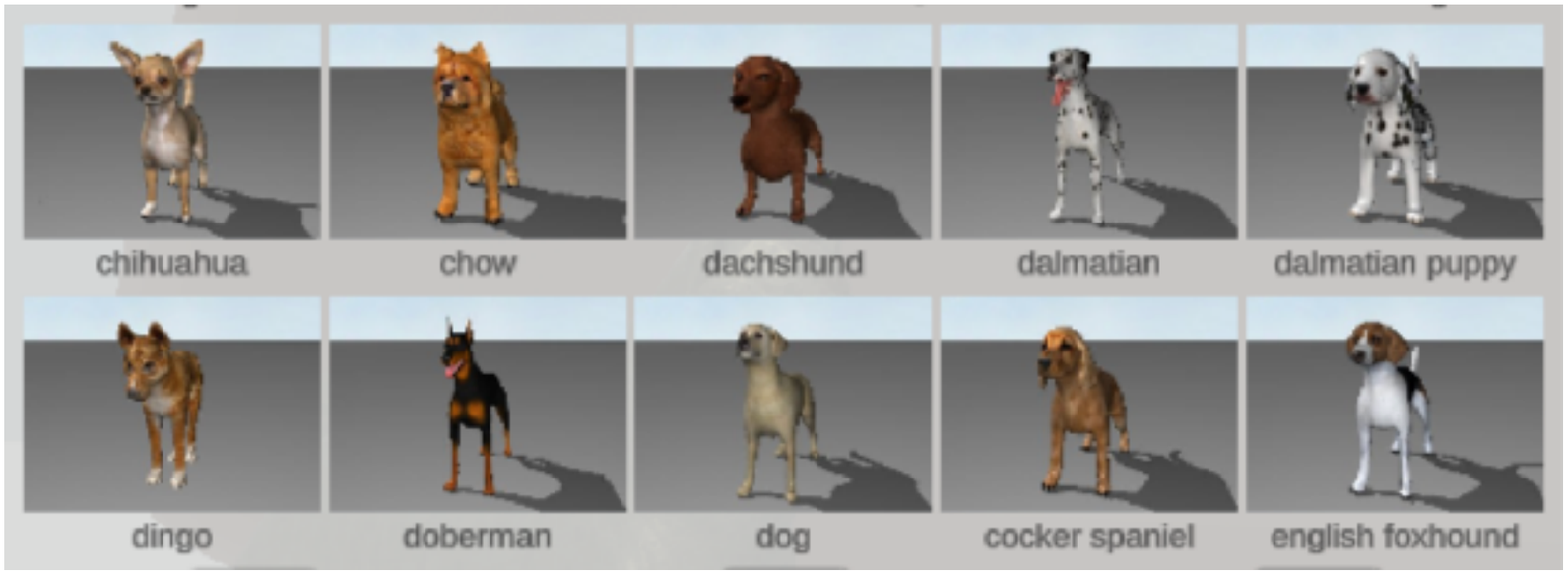


Multiple inheritance for a pirate zombie object



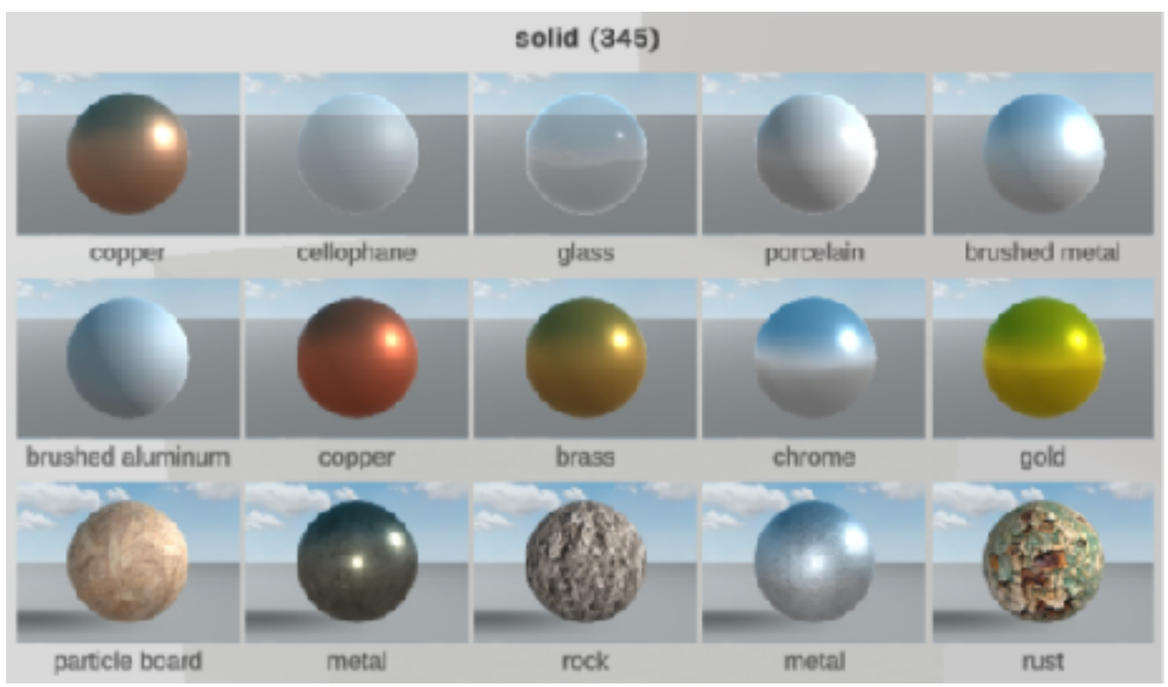
Content types: 3D objects, images, materials, audio samples

12K 3D objects with 70K geometric parts

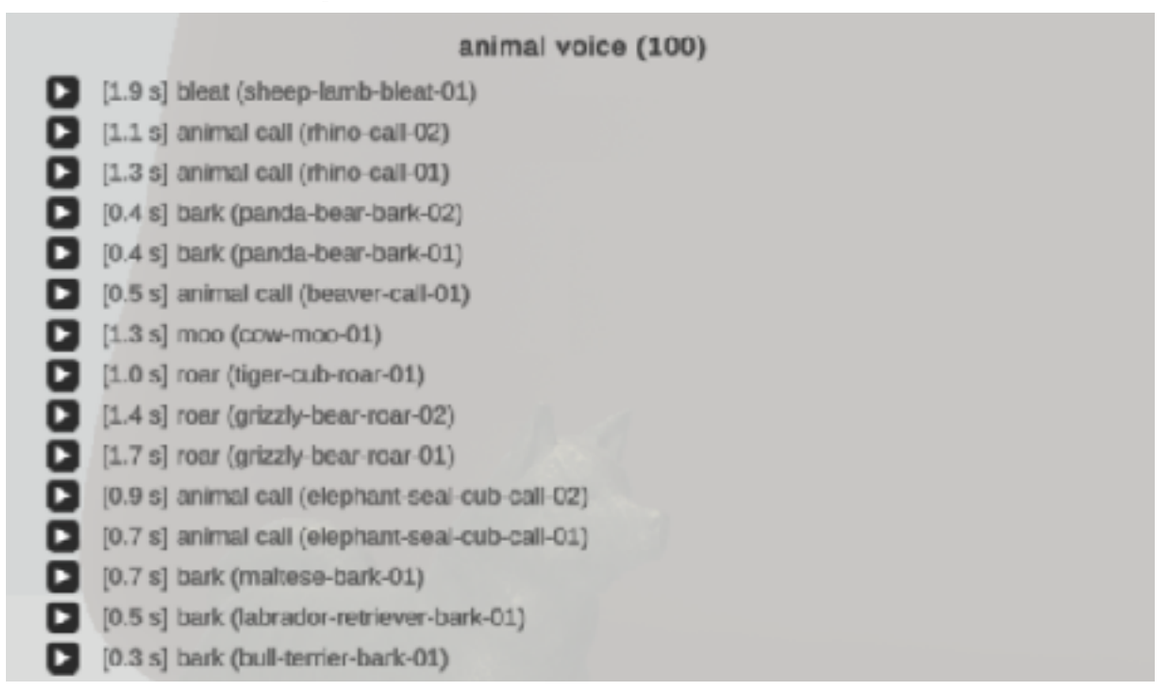


100K images

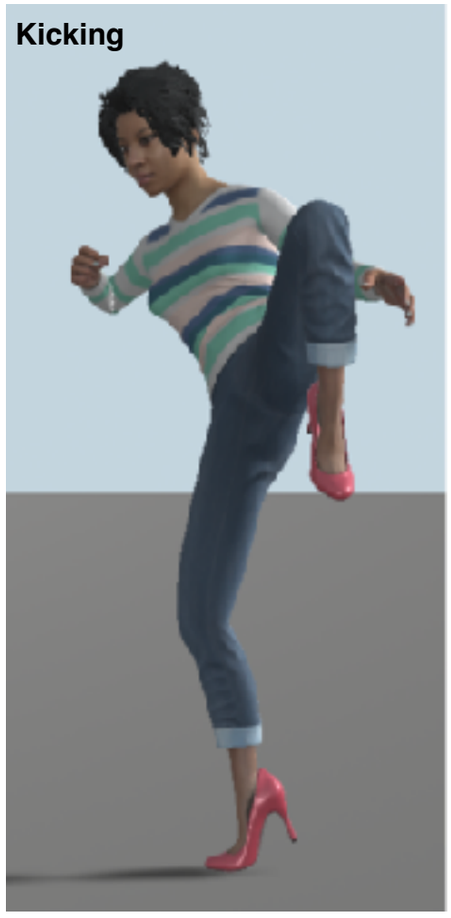
600 materials



14K Audio samples



Posed characters

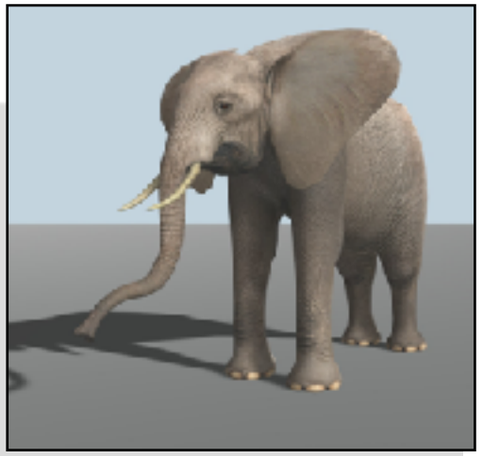


Lexicon & Concepts → Objects

15K nouns

ELECTRONIC-CAR-KEY.N (electronic car key, car key)
ELECTRONIC-CIGARETTE.N (electronic cigarette)
ELECTRONIC-DEVICE.N (electronic device, electronics)
ELECTRONIC-FUSE.N (electronic fuse, fuse)
ELECTRONIC-INDUCTOR.N (electronic inductor, inductor)
ELECTRONIC-ORGAN.N (electronic organ)
ELECTRONIC-PORT.N (electronic port)
ELEGY.N (elegy)
ELEMENT.N (element)
ELEMENTARY-SCHOOL-STUDENT.N (elementary school student)
ELENA-NAME.N (Elena)
ELEONOR-NAME.N (Eleonor)
ELEPHANT.N (elephant)
ELEPHANT-CALF.N (elephant calf)
EMOJI-ELEPHANT.N (elephant emoji)
ELEPHANT-GRAY-COLOR.N (elephant gray)
ELEPHANT-SEAL.N (elephant seal)
ELEPHANT-TRUNK.N (elephant trunk)
ELEVATION.N (elevation)
ELEVATOR.N (elevator)
ELEVATOR-BELL.N (elevator bell)
ELEVATOR-DOOR.N (elevator door)
EMOJI-ELEVATOR.N (elevator emoji)
EMOJI-ELEVEN_OCLOCK.N (eleven o'clock emoji)
EMOJI-ELEVEN_THIRTY.N (eleven-thirty emoji)
ELF.N (elf)
EMOJI-ELF.N (elf emoji)
ELFA-NAME.N (Elfa)
ELFREDA-NAME.N (Elfreda)
ELI-MANNING.N (eli manning, manning, eli, eli manning)
ELIANE-NAME.N (Eliane)
ELIAS-NAME.N (Elias)
BIBLICAL-ELIJAH.N (elijah)
ELINOR-NAME.N (Elinor)
ELIOT-NAME.N (Eliot)
ELISABETH-NAME.N (Elisabeth)
ELISE-NAME.N (Elise)

Objects are leaf nodes in the concept ontology



elephant.n (*elephant*)

Supernodes:
pachyderm.n *pachyderm* (5)

Subnodes:
african-elephant.n *african elephant* (1)
asian-elephant.n *asian elephant* (2)
obj-african_elephant-hm0181 *obj-hm0181, hm0181, obj-african_elephant-hm0181*

[Indirect] african-elephant.n
obj-asian_elephant-hm0190 *obj-hm0190, hm0190, obj-asian_elephant-hm0190*

[Indirect] asian-elephant.n
obj-cartoon-elephant-ind0975 *obj-ind0975, ind0975, obj-cartoon-elephant-ind0975*

obj-alien-circus-elephant-ind1485 *obj-ind1485, ind1485, obj-alien-circus-elephant-ind1485*

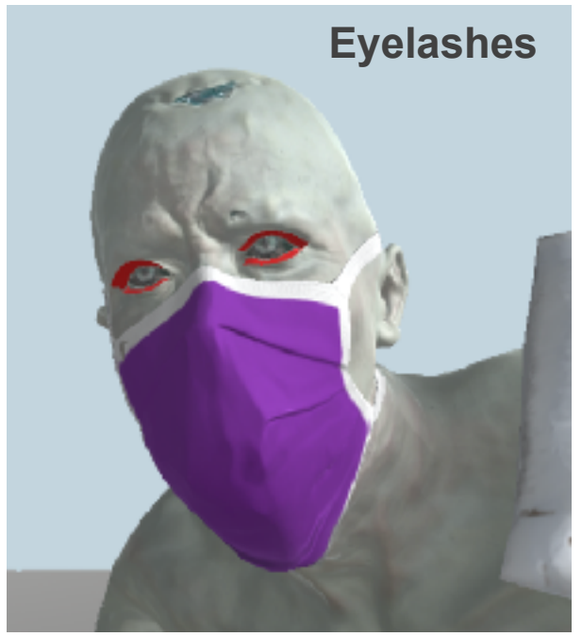
obj-elephant-swimming-ind1583 *obj-ind1583, ind1583, obj-elephant-swimming-ind1583*

obj-elephant-vp1336 *obj-vp1336, vp1336, obj-elephant-vp1336*


3D Object parts

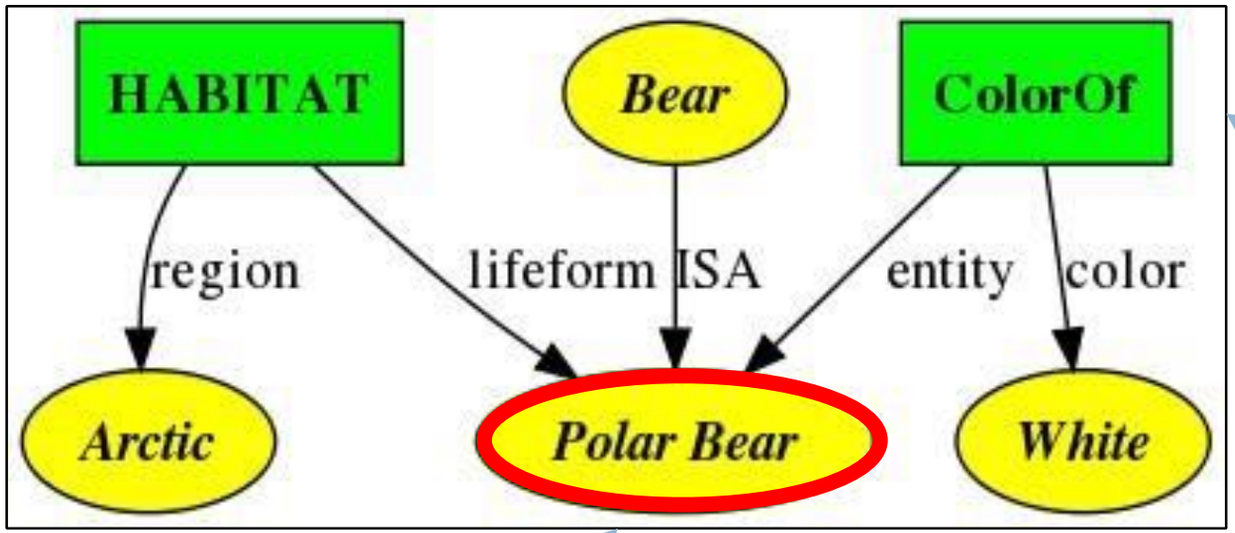
Polygonal 3D objects have geometrically-defined parts. This is done very idiosyncratically.

This object has 5 parts that can be referenced with language. E.g. “The zombie’s mask is blue”



Asserted relations → world knowledge and word meaning

polar bear: a white colored bear that lives in arctic

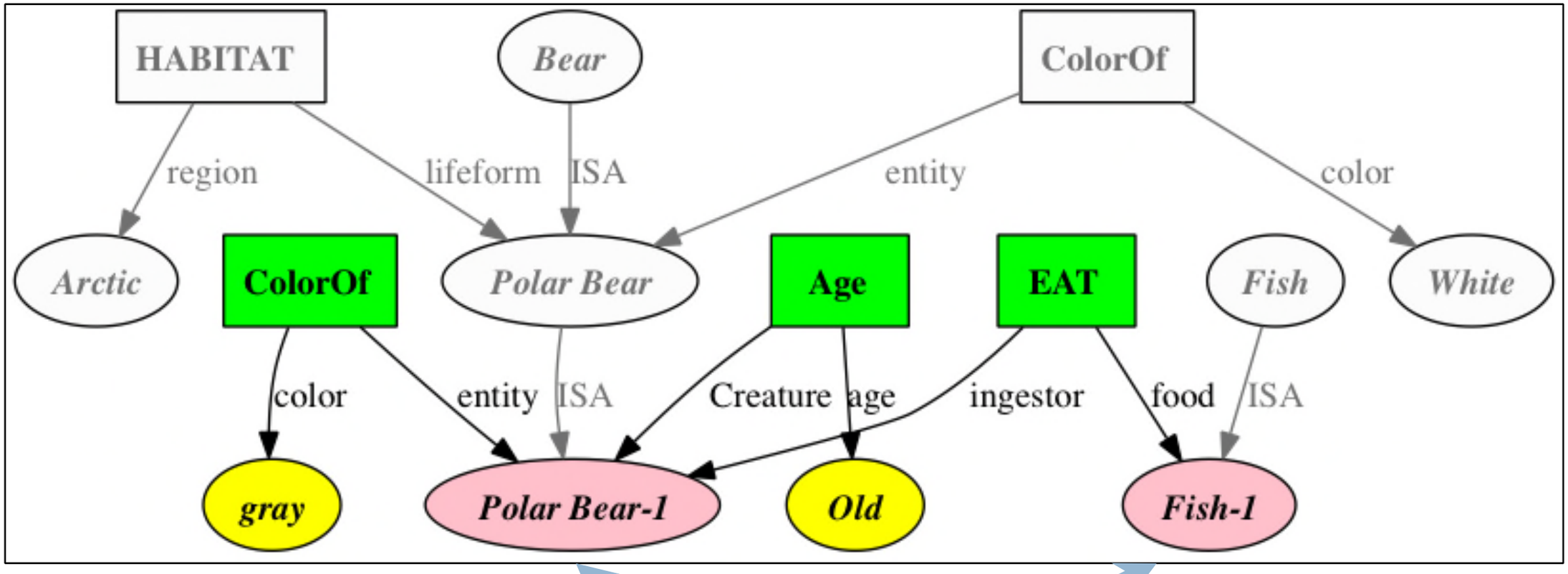


Semantic relation gives meaning to polar bear concept

Concept for "polar bear"

…and sentence meaning

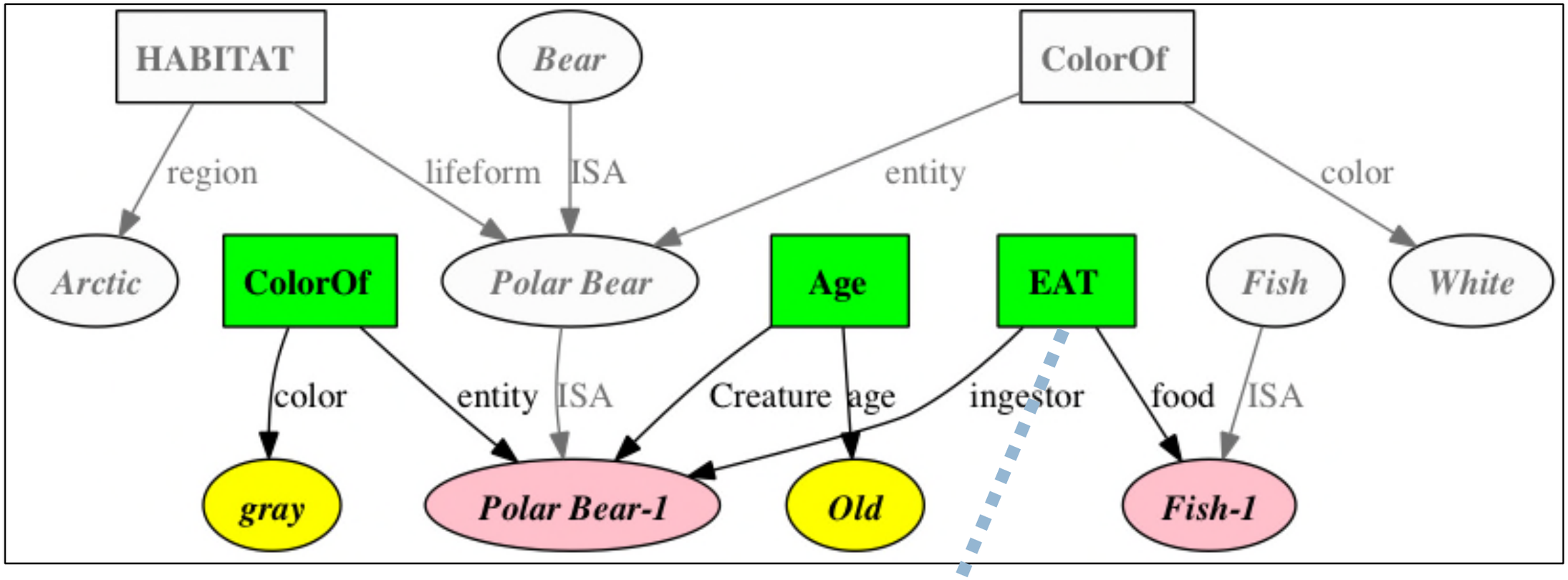
The old gray polar bear ate the fish



Concepts for a particular polar bear and fish

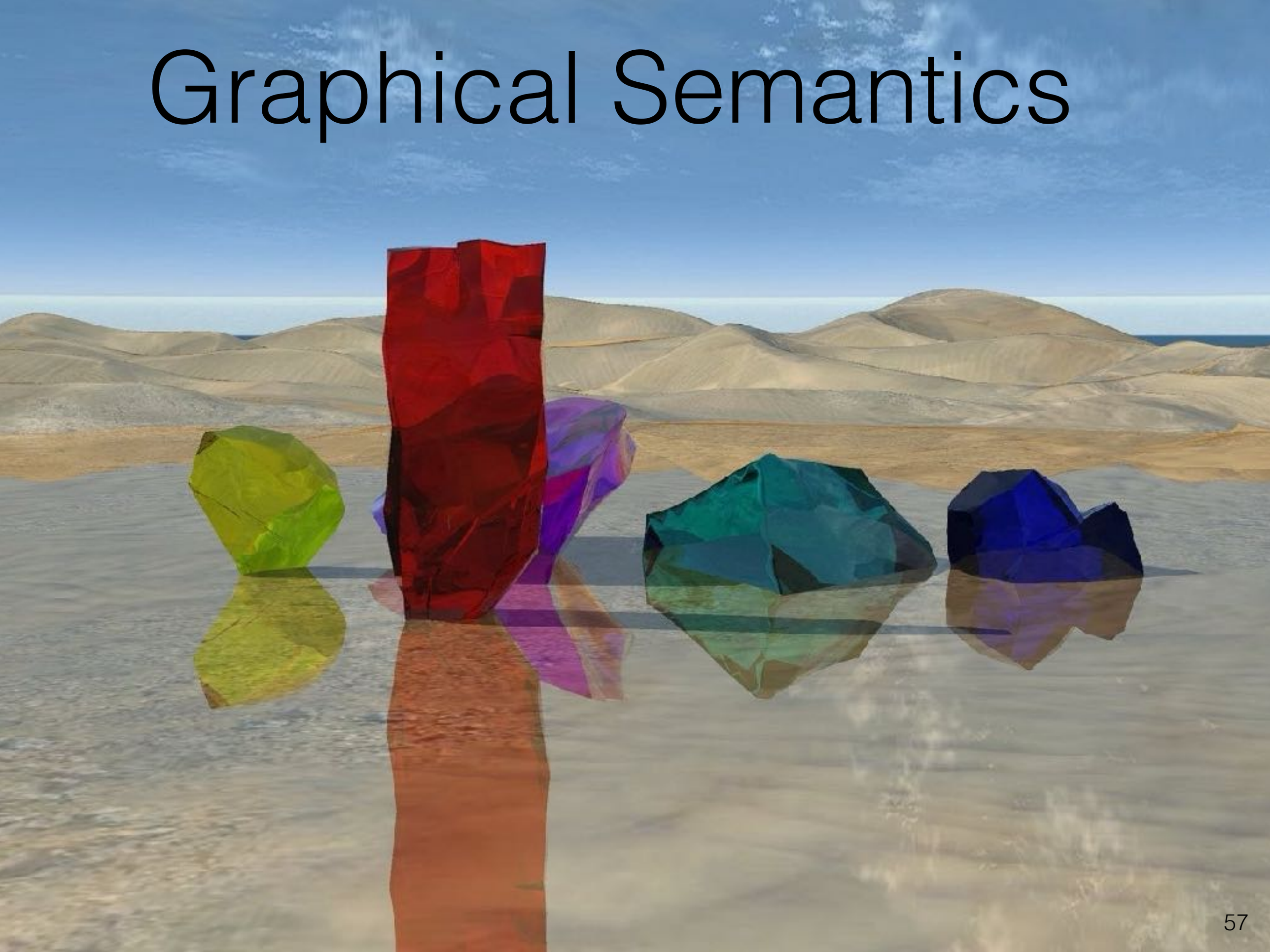
…grounding Sentence Meaning

The old gray polar bear ate the fish



Decompose *eat* into a vignette and then to graphical relations

Graphical Semantics



Graphical Semantics

What is needed to specify the graphical structure of scenes?

- Graphical properties on objects
- Graphical primitives (constraints) to compose the scene

The boat is <u>in</u> the ocean. The dog is <u>in</u> the boat.

- Resolve "in" to more specific spatial relations using object semantics
  - *Boat in water* ➔ EMBEDDED-IN
  - *Dog in boat* ➔ IN-CUPPED-REGION
- Depends on object shape and function

Intrinsic Properties of Objects

Represented as asserted semantic relations

All 3D objects have an default ***size*** and ***orientation*** that can be inherited through ISA hierarchy

Other graphical and functional properties. E.g.

- Constrained axes, segmentation, length axis
- Embedding distance (e.g. boat waterline)
- Vertical surface item (e.g. sconce)

Spatial regions as affordances

Dominant parts for changing color

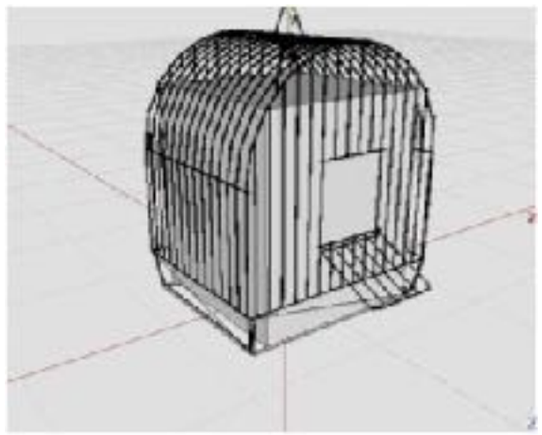
Spatial Regions as Affordances

BASE

CUP

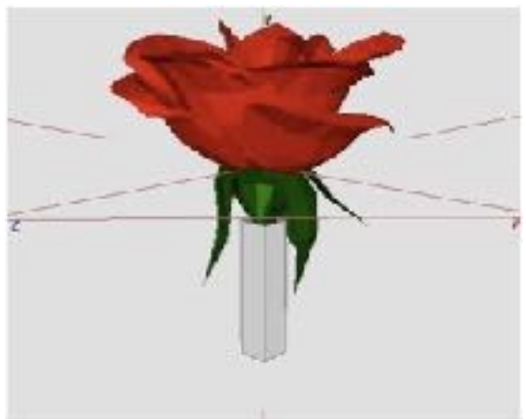
ON

CANOPY



ENCLOSURE

HANDLE

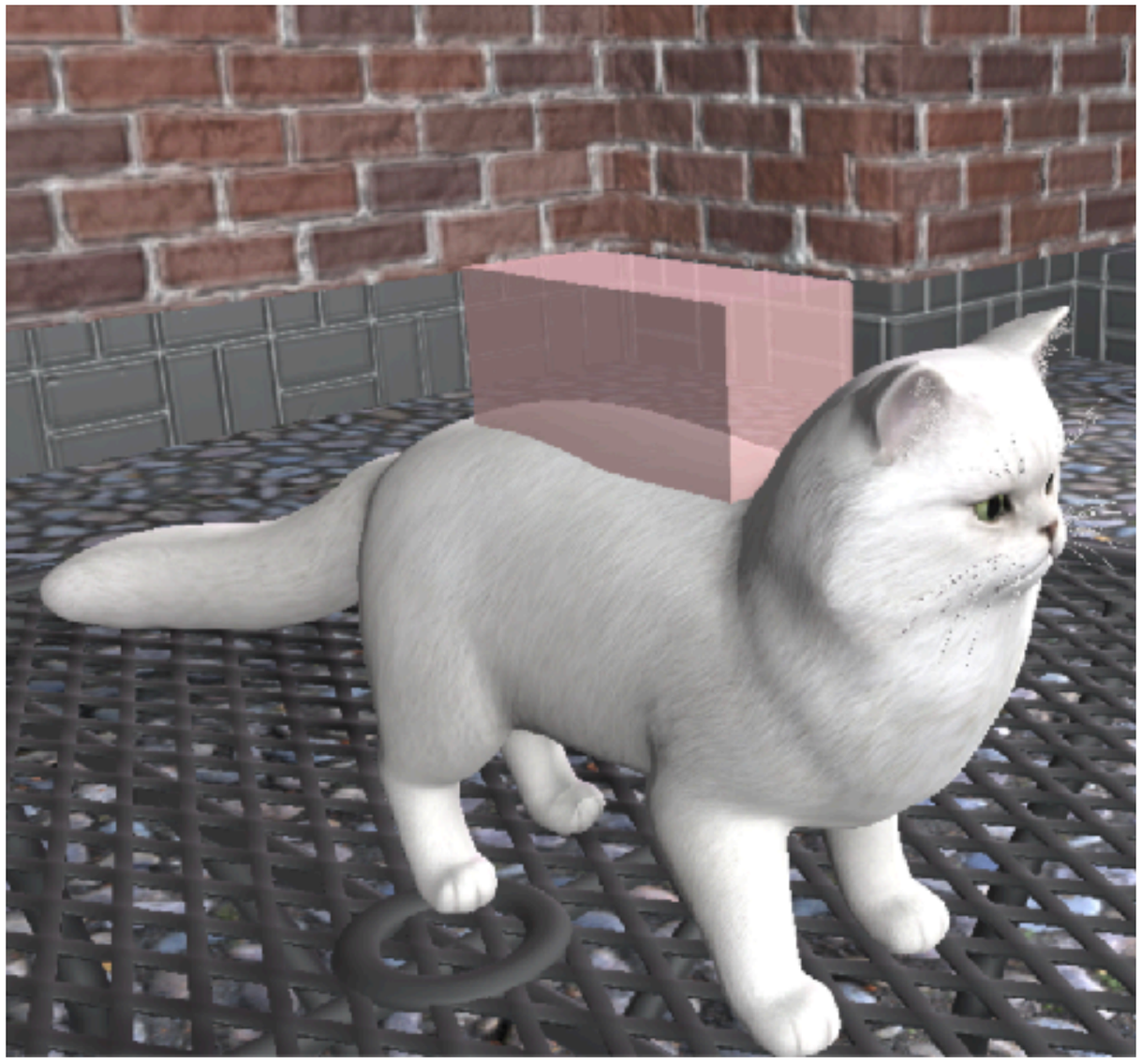


STEM

WALL

https://www.cs.columbia.edu/speech/PaperFiles/2001/wordseye_siggraph.pdf

Spatial prepositions



Salient regions are pre-designated on a per-object basis. e.g. for ON the cat

"A robin is on the cat"

Interpretations of spatial prepositions

Spatial relation	Scene elements
ENCLOSED-IN	*Chicken in cage*
EMBEDDED-IN	*Horse in ground*
IN-CUP	*Chicken in bowl*
ON-TOP-SURFACE	*Apple on wall*
ON-VERTICAL-SURFACE	*Picture on wall*
PATTERN-ON	*Brick-texture on wall*
UNDER-CANOPY	*Vase under umbrella*
UNDER-BASE	*Rug under table*
STEM-IN-CUP	*Flower in vase*
LATERALLY RELATED	*Wall behind table*
LENGTH AXIS	*Wall*
DEFAULT SIZE/DIRECTION	All objects
REGION	*Right side of*
DISTANCE	*2 feet behind*
SIZE	*Small* and *16 ft*
ORIENTATION	*facing*

Input text: *A large magenta flower is in a small vase. The vase is under an umbrella. The umbrella is on the right side of a table. A picture of a woman is on the left side of a 16 foot long wall. A brick texture is on the wall. The wall is 2 feet behind the table. A small brown horse is in the ground. It is a foot to the left of the table. A red chicken is in a birdcage. The cage is to the right of the table. A huge apple is on the wall. It is to the left of the picture. A large rug is under the table. A small blue chicken is in a large flower cereal bowl. A pink mouse is on a small chair. The chair is 5 inches to the left of the bowl. The bowl is in front of the table. The red chicken is facing the blue chicken. . .*

Color parts

Only the dominant “color” parts should change when we specify a new color for an object.

The dog

The dog is blue

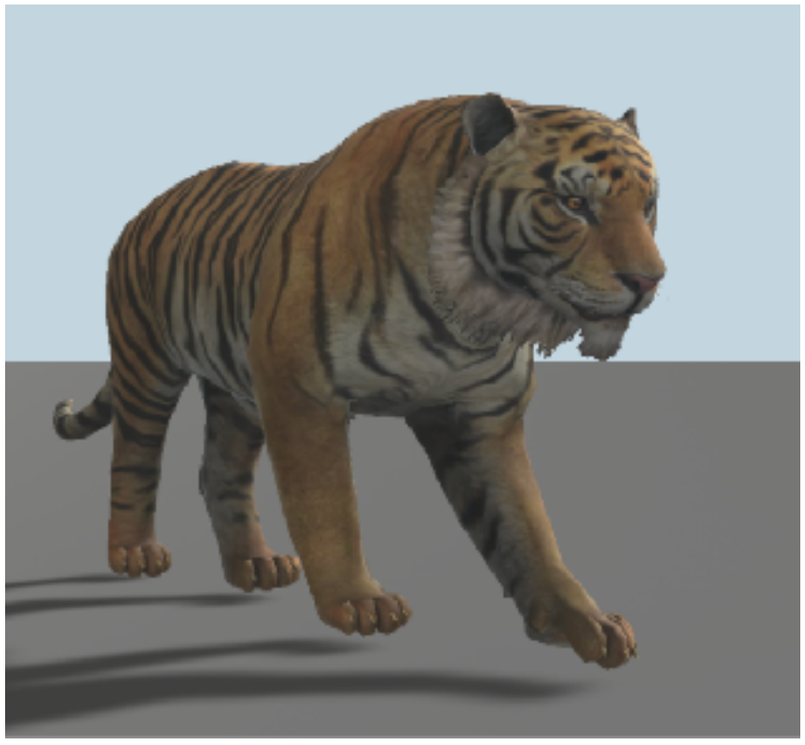
Change just the dominant color parts by default.

The dog is entirely blue

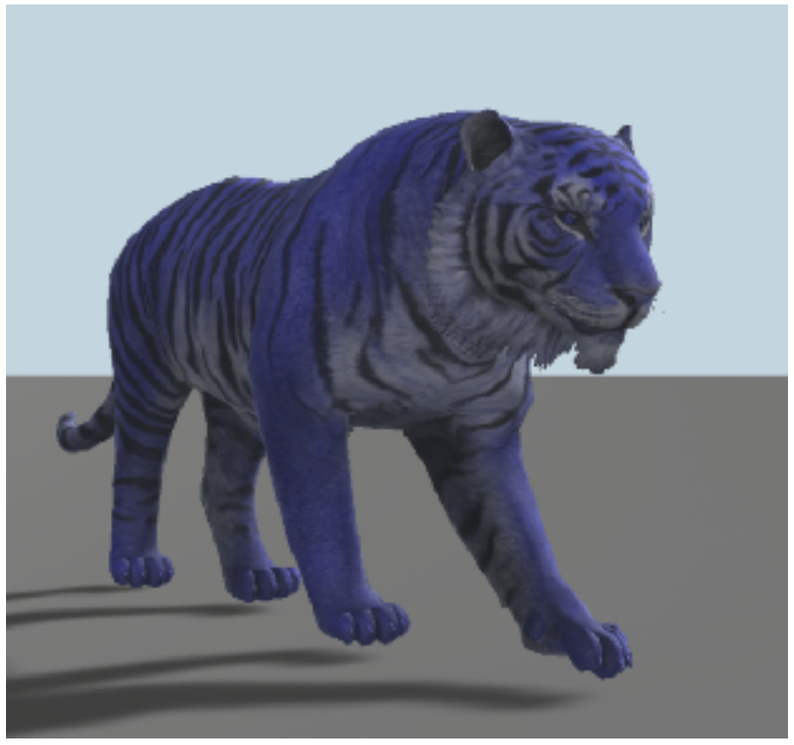
Force all parts to change color.

Colors on textures

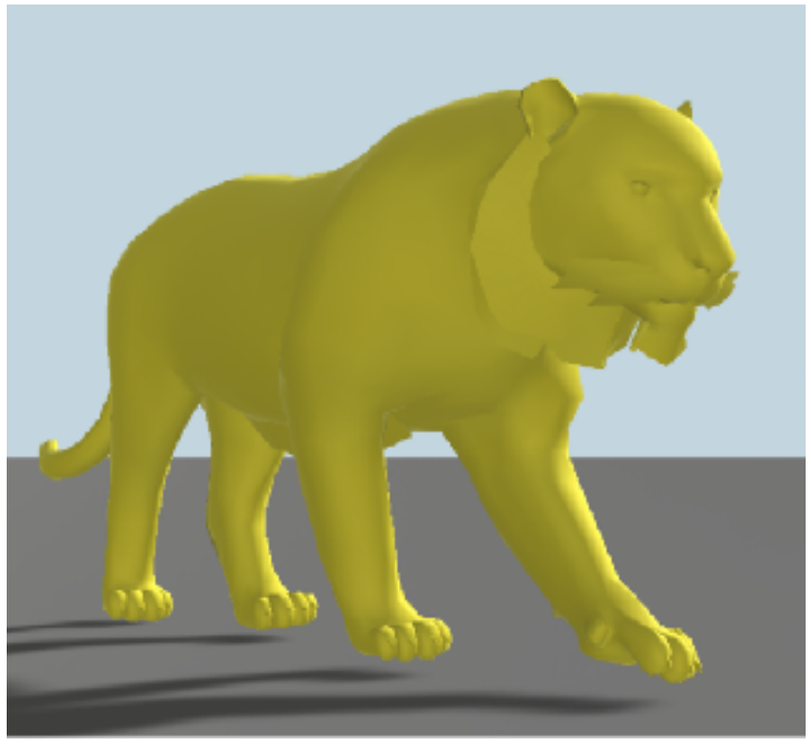
This object has only one part and is textured. So the options are to tint the entire texture or replace it with a solid color.



The tiger



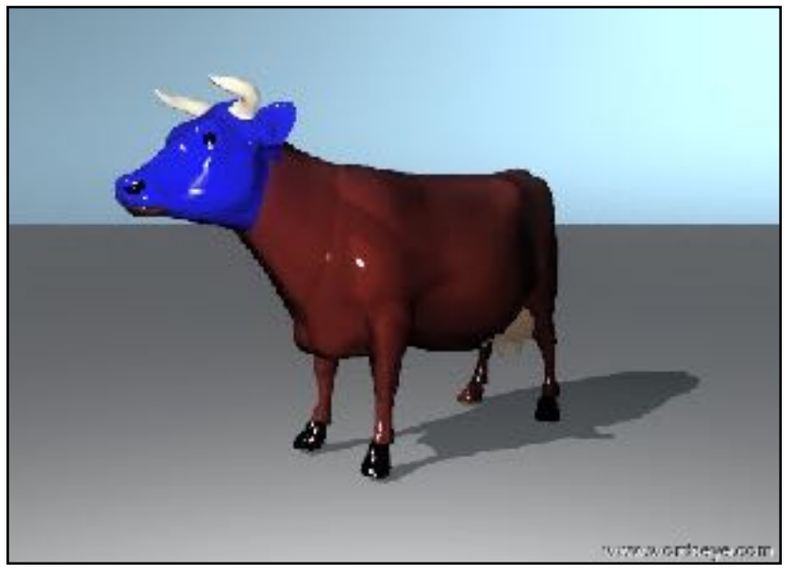
The tiger is blue



The tiger is solid yellow

Interpretations of “of”

Containment: *bowl of cats*



Part: *head of the cow*

Dimension: *height of horse is..*

Grouping: *stack of cats*

Substance: *horse of stone*

Representation: *Picture of girl*

Resolving "Of" based on arguments

Text (A of B)	Resulting Semantic Relation	Conditions
Bowl of cherries	CONTAINER-OF (bowl, cherries)	A=container, B=plurality-or-mass
Slab of concrete	SUBSTANCE-OF (slab, concrete)	A=entity, B=substance
Picture of the girl	REPRESENTS (picture, girl)	A=representing-entity, B=entity
Arm of the chair	PART-OF (chair, arm)	A=part-of(B), B=entity
Height of the tree	DIMENSION-OF (height, tree)	A=size-property, B=physical-entity
Stack of plates	GROUPING-OF (stack,plates)	A=arrangement, B=plurality

Interpreted and inferred semantic relations

The goal is to resolve syntactic constructs to semantic relations. Some cases:

Prepositions (interpret/resolve the relation)

- *the picture of flowers* → *the picture* ***represents*** *flowers*

  arg1 preposition arg2 → representation-of(**arg1, arg2**)

Noun-noun compounds (infer the missing relation)

- *the wood table* → *the table is* ***made of*** *wood*

  arg1 arg2 → made-of(**arg1, arg2**)

Metonymy and regular polysemy (infer the missing relation and arg)

- *the cereal on the shelf* → *the* ***cereal box*** *is on shelf … and it* ***contains*** *cereal*

  arg → container-of(cereal-box, **arg**)

Spatial reference frames

Spatial reference frames

Natural language descriptions of spatial scenes describe the location of one thing with respect to other things.

In a spatial description, something (the ***figure***) is generally located with respect to something else (the ***ground***).

We want to know in which direction from a ground we need to search to find the figure. A ***coordinate system*** comes into play.

Levinson/Wilkins: The background to the study of the language of space. Chapter 1 in Grammars of Space — Explorations in Cognitive Diversity

https://pure.mpg.de/rest/items/item_59541_3/component/file_59542/content

Spatial reference frames

Different types of reference frames:

Egocentric — coordinate systems anchored to the body of an *observer*

Allocentric — coordinate systems defined by *intrinsic* features of objects in the environment, independent of the observer.

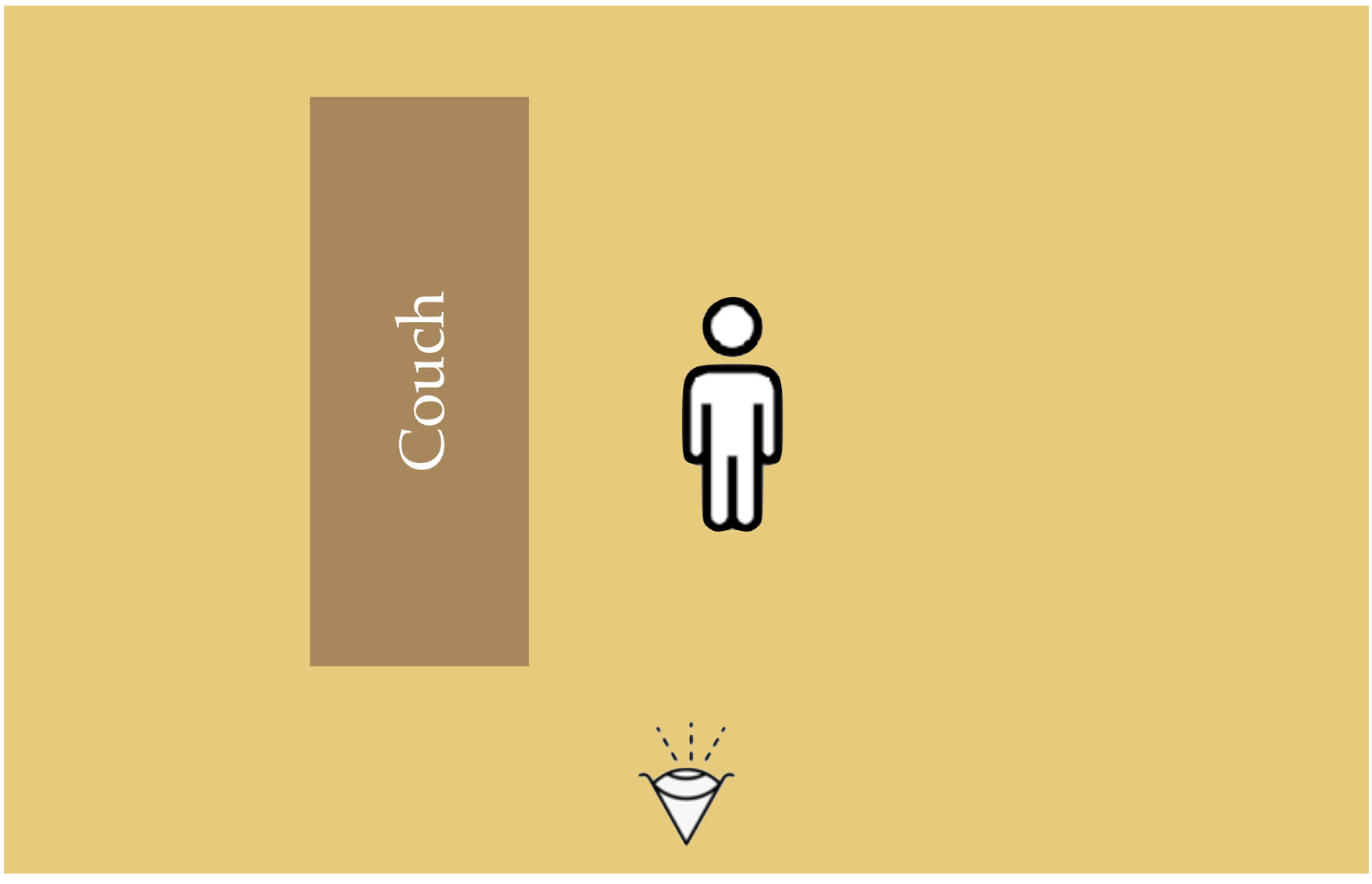
- **Object-centric:** Relative to the *ground* object (in a figure/ground relation)
- **Stage-centric:** Relative to an implicit environment

Absolute — e.g. Boston is *north* and *east* of New York.

Combinations — where egocentric position or an additional context mediates an allocentric reference frame

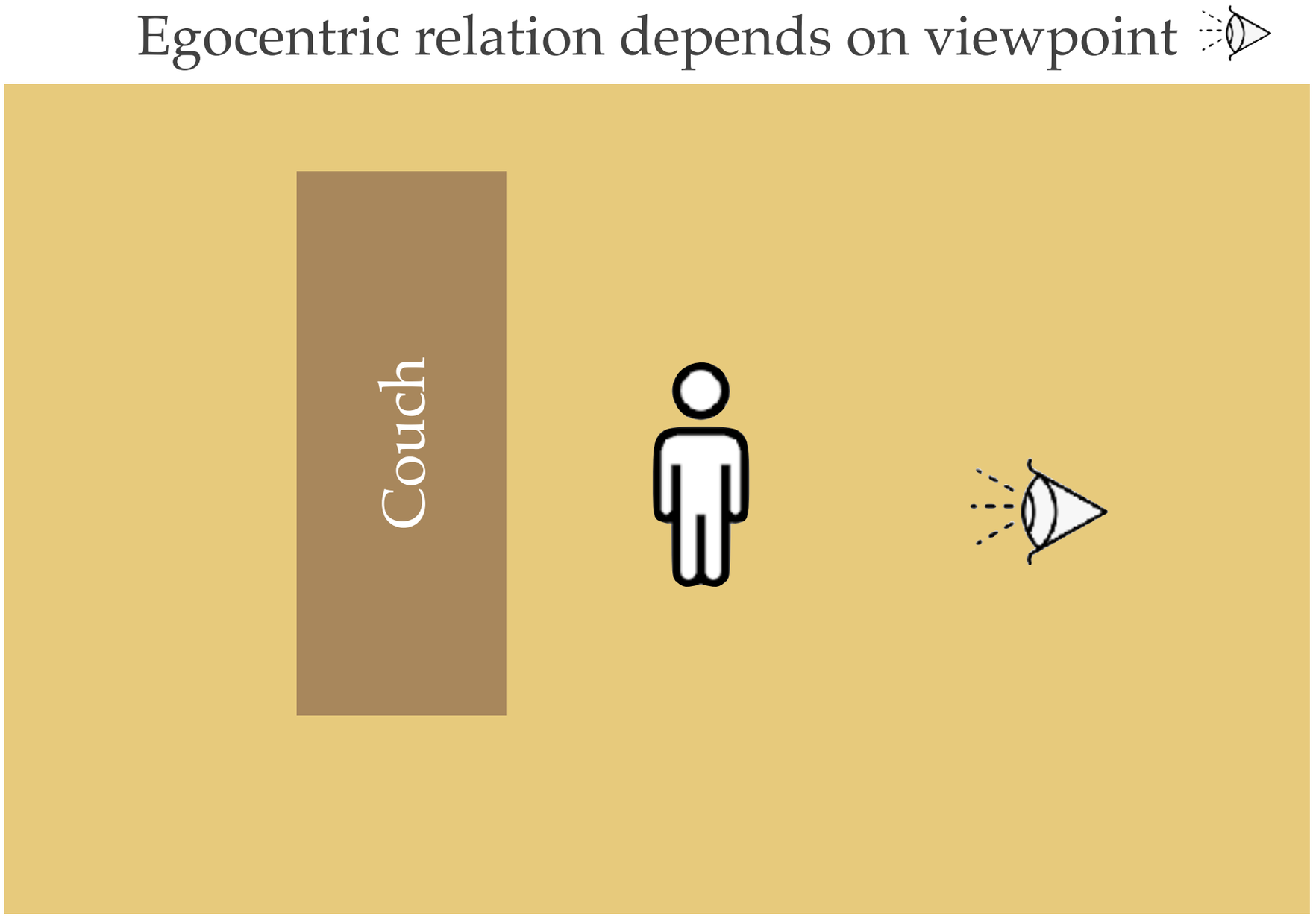
Egocentric reference frame

Egocentric relation depends on viewpoint



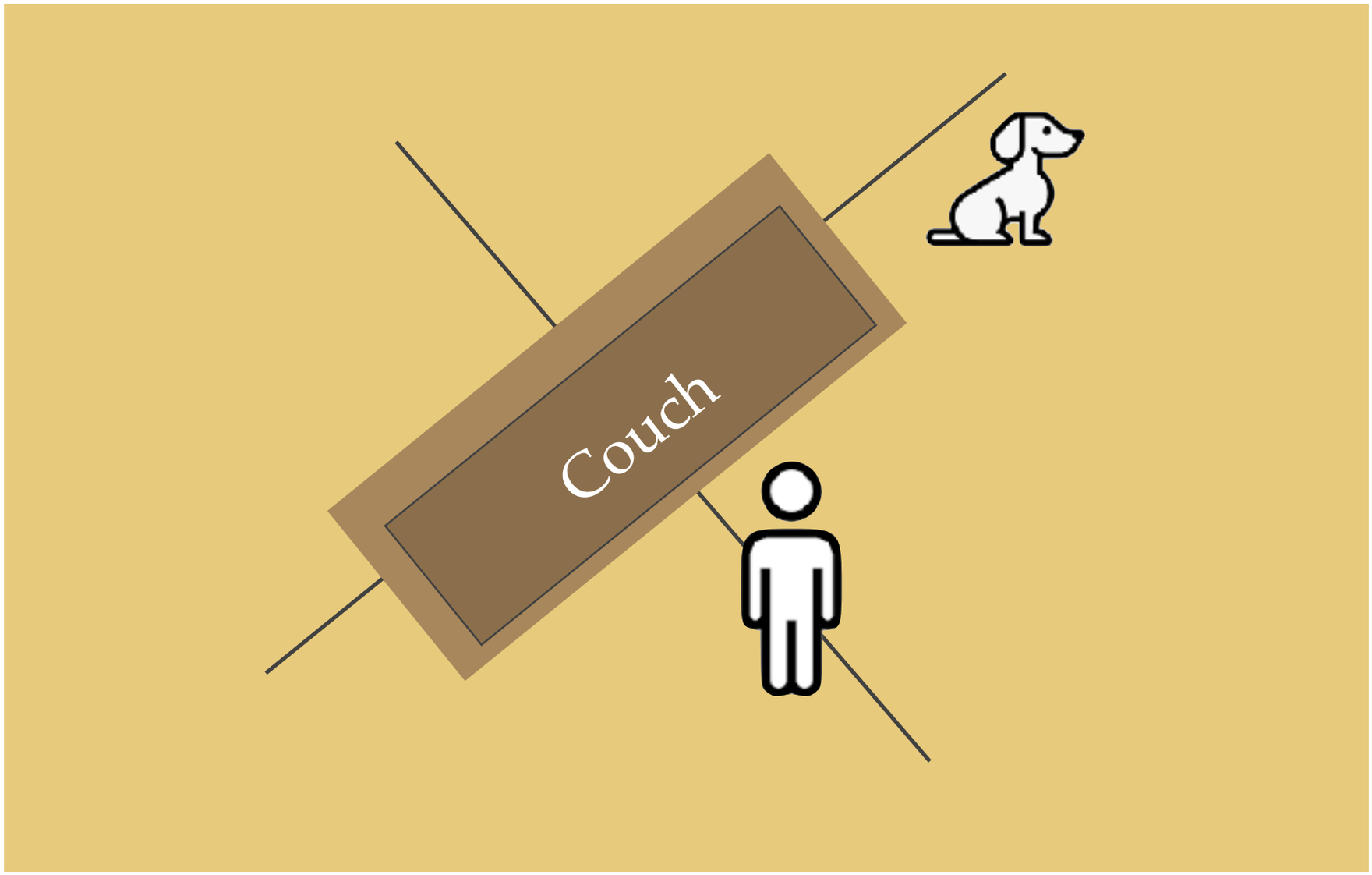
*The man is **right of** the couch*

Egocentric reference frame



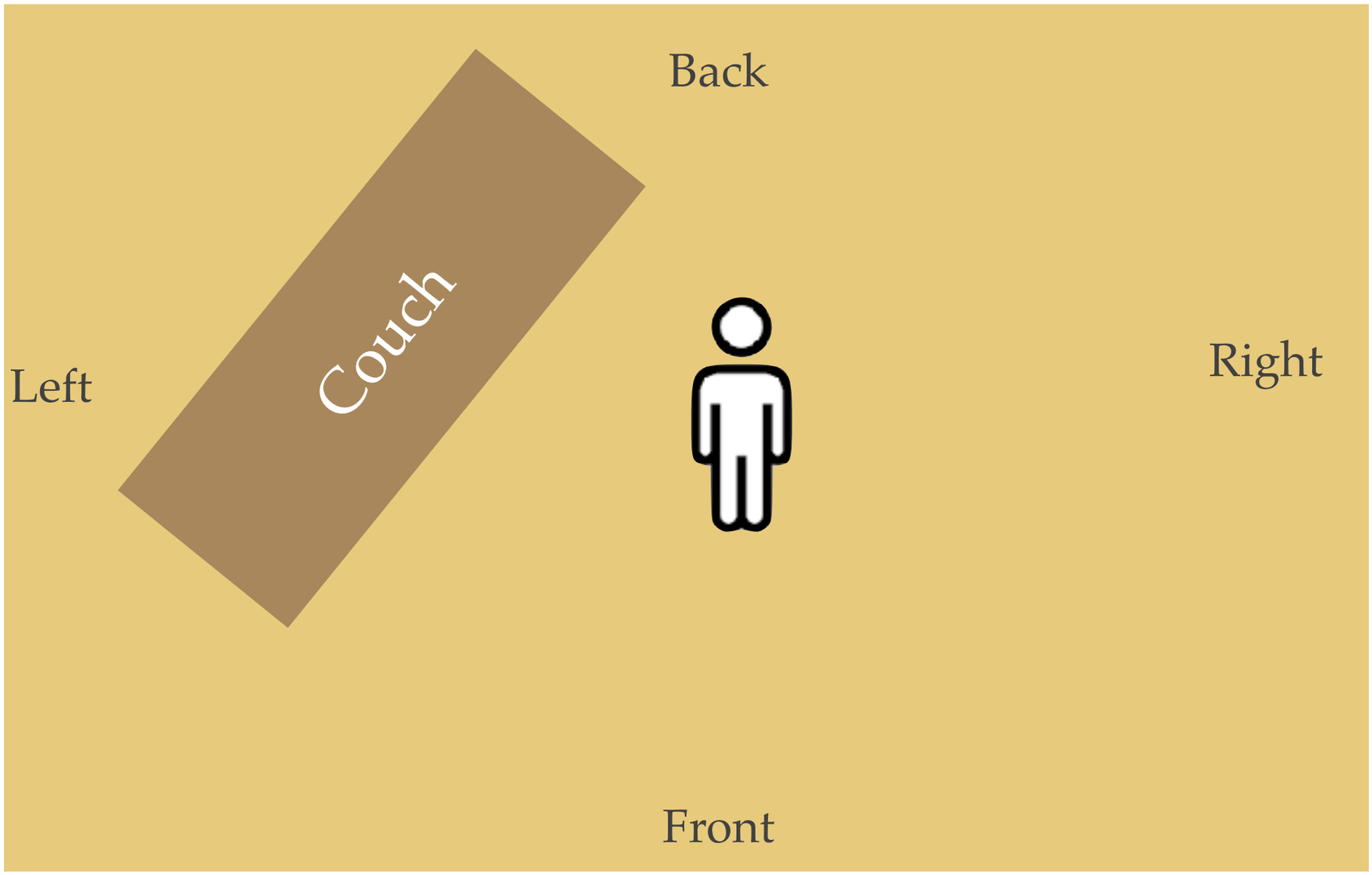
*The man is **in front of** the couch*

Object-centric (couch)



*The man is **in front of** the couch.*
*The dog is **right of** the couch.*

Stage-centric (room)



*The man **is right of** the couch*

Frames of reference (egocentric)

Same scene — Description changes for same scene with different **egocentric** reference frames

Egocentric: *The zombie is **left of** the tree.*

Egocentric: *The zombie is **right of** of the tree.*
(Looking from the house)

Frames of reference (object-centric)

Orientation of car establishes **object-centric** reference frame

Object-centric (car): *The zombie is **in front of** the car*

Ego-centric/stage-centric: *The zombie is **right of** the car*

Object-centric (car): *The zombie is **behind** the car*

Ego-centric/stage-centric: *The zombie is **left of** the car*

Frames of reference (combo)

Same scene — Fence has no well-defined front vs back. So, **object-centric** reference frame is **mediated by viewpoint**.

Object-centric (fence): *The zombie is **behind** the fence*

Object-centric (fence): *The zombie is **in front of** the fence*

Frames of reference (stage-centric)

The vase has **no intrinsic orientation**. So use egocentric or stage-centric (house)

Stage-centric: *The zombie is **right of** the vase*

Stage-centric: *The zombie is **left of** the vase*

Object-centric vs stage-centric

Stage-centric: *An astronaut is **right of** the couch*
Stage-centric: *A dog is **in front of** the couch*

Object-centric (couch): *A astronaut is **in front of** the couch*
Object-centric (couch): *A dog is **left of** the couch*

Frames of reference (some factors)

What is the most natural frame of reference to use, given:

- Input text
- Existing scene

Factors

- Do surrounding objects imply or establish a stage-centric reference frame?
- Does viewer (ego-centric) location mediate the choice of object-centric ref frame?
- *Ground* object intrinsic properties: have an identifiable front vs back? Does it have any natural orientation or is it the same from all angles.
- Does the text imply object-centric or stage-centric or viewer centric
- Does figure orientation matter?
- Do additional (possibly pre-existing) constraints affect the interpretation of the given constraint.
- Is the model simple enough to allow the user predict the system's interpretation.

Referring expressions

Reference resolution

Given versus ***new*** information is a distinction between information that is assumed or supplied by the speaker and that which is presented for the first time. [Prince 1981]

Textual co-references: In text, we must determine if two references, within the text, are to the same object or not. If they are, then they are represented by a single entity, and any references to them (including attributes) are merged.

Text referring to scene entities: Text can also refer to objects in the current scene. References must then be merged with the scene. Otherwise a new object is introduced into the scene.

ftp://babel.ling.upenn.edu/papers/faculty/ellen_prince/givennew.pdf

Reference resolution (within text)

Definite vs indefinite articles.

- "*A chair is near the couch. A cat is on the chair*" → [merge]
- "*A chair is near the couch. A cat is on a chair*" → [new]

Hypernyms and hyponyms

- "*A Poodle is sleeping on the lawn. The dog is very old.*" → [merge]
- "*A dog and cat were fighting. The poodle is near a tree.*" → [new]

Membership in collections

- "*A cat and dog are at the door. The animals are hungry.*" → [merge]
- "Five cats and three dogs are out in the yard. A dog started chasing a cat." → [merge new individuals with collection]

Attributes as identifiers

- "*The red tree is 20 feet tall. The red tree is near the house*" → [merge]
- "*The red tree is 20 feet tall. The tall tree is near the house*" → [new]

Various other factors (e.g. pragmatic considerations; position in sentence — subject or direct object)

Reference resolution (merging with scene)

Many of the same factors (indefinite/definite reference, subtypes, …) apply. But the objects in a scene are fully grounded. That allows arbitrary (previously unstated) properties of those objects to be referenced and cause merging.

Definite/indefinite references

- Chair in scene + "*The chair*…" → [merge if found]
- Chair in scene + "*A chair*…" → [new]

Intrinsic properties

- "*The dead tree is*…" → [merge if found]

Ordinals

- "The *1st tree* is red. The *2nd tree* is blue" → [find and merge separately]

Referenced properties vs assigned properties

- "*The dog near the tree is big*" → [find and merge from position… and then assign size]
- "*The big dog is near the tree*" → [find and merge from size…and then assign position]

Relational attributes (intrinsic or computed)

- "*The man with the hat is*…" → [find and merge]

Merging with scene and text

Initial scene

Add new objects textually

The man is left of the tree.
The hunter is right of the tree.

- No textual merge between man and hunter.
- No scene merge for man or hunter, so new objects added.
- Tree merges textually with itself and then with the tree in the scene.

Refer and modify objects in scene

Santa is blue.
He is facing the woman

- Santa and He textually merge. (Since Santa is a man).
- Santa/He merge with the scene object.
- Woman merges with the scene object.

Note: Original "*Hunter*" happened to be a woman and original "*man*" happened to be Santa. They are then referred to based on the resulting scene objects .

Computed scene references

The man near the table is facing left.

Initial scene

The short man is tinted blue.

Gestural references

Deictic gestures

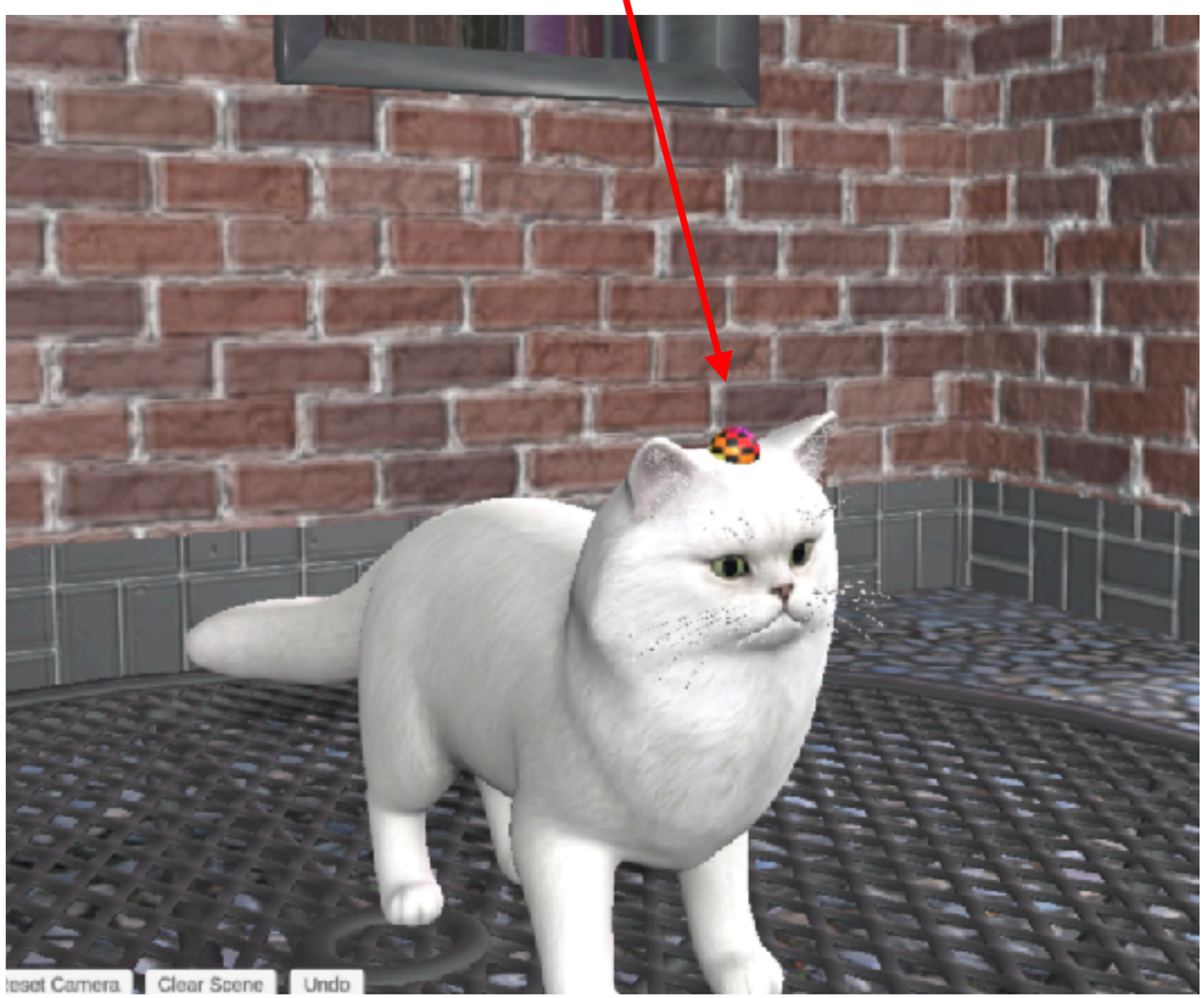
These gestures are also known as pointing where children extend their index finger, although any other body part could also be used, to single out an object of interest. Deictic gestures occur across cultures and indicate that infants are aware of what other people pay attention to.

Gestures in language acquisition - Wikipedia

Examples: (indicating) *here, there, that, it, those*

Referring to HERE and THERE

Point at a location



"A small robin is THERE"

Future work

- Deformable objects (e.g. clothing, rope, food, …)
- Non-discrete objects (e.g. hills and valleys and holes that blend into each other)
- Environmental effects (snow, rain, smoke,…)
- Computational geometry to automatically designate and modify object parts and/or to generated object variants (e.g. a chair with 20 inch tall legs)
- Dynamically posable characters
- Character animation and facial expressions

Many more…

End