How New Tasks and Datasets Have Enabled Progress in NLP

Julia Hockenmaier
University of Illinois
juliahmr@illinois.edu

IJCNLP-AACL 2023 November 2 2023 Bali, Indonesia

What is the current state of NLP?

Web-scale LLMs work *amazingly* well for many tasks

But they require huge amounts of money and resources
(very large teams, many large GPUs)

These LLMs are created by a multi-billion dollar industry

A few companies have more money than any country's funding agencies

Why should you listen to *my* talk?

No new LLMs, no bigger, better models

But you'll see some "emergent abilities" of VSLMs
(Very Small Language Models)

Just a few examples of small-scale efforts to get computers to understand and produce (some aspects of) language

How do we made progress in NLP?

We push the performance of NLP through…
… better representations,
… better models,
… better algorithms

We push the scope of NLP through…
… new datasets
… new tasks

What does it take to *understand* Language?

What does it take to *'understand'* language?

People are shopping groceries
in a supermarket

People are **shopping** groceries **in a supermarket**

People are shopping **groceries in a supermarket**

Natural language understanding involves the ability to **resolve (syntactic) ambiguities**

How do you get a computer to resolve (syntactic) ambiguities?

Statistical Parsing

Grammar
Defines the sentences of the language and their possible structures (trees τ)

Probability model
Assigns a score P(τ) to each tree τ

Parsing algorithm
Returns the best tree τ* = argmax P(τ) for each sentence

Statistical Parsing

Treebank
(labeled training data)

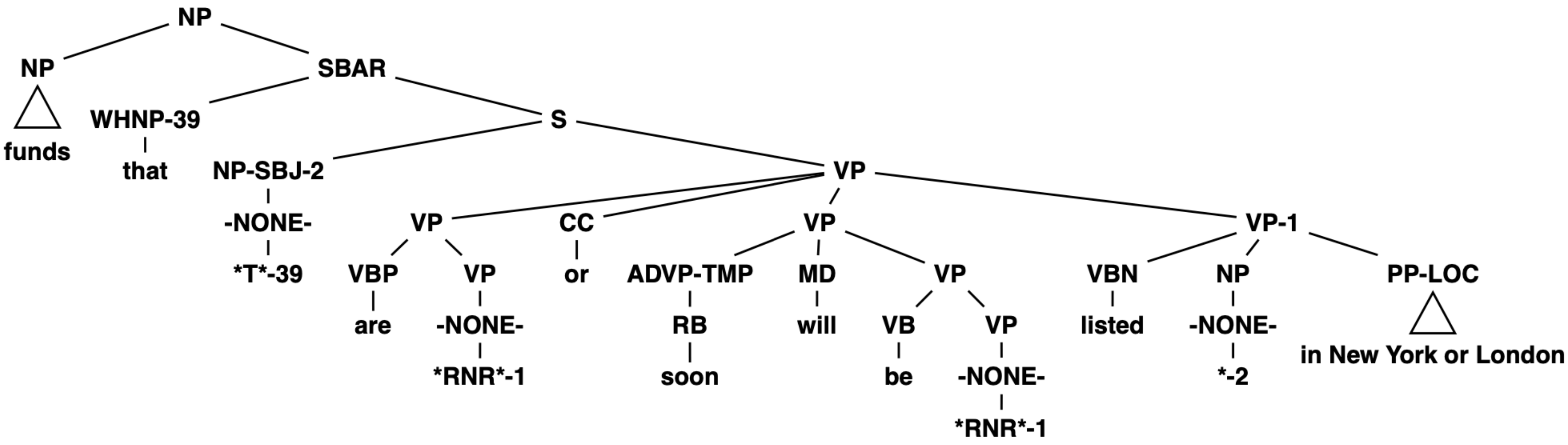
Grammar
Defines the sentences of the language and their possible structures (trees τ)

Probability model
Assigns a score P(τ) to each tree τ

Parsing algorithm
Returns the best tree
τ* = argmax P(τ)
for each sentence

Penn Treebank
(Marcus et al., 1994)

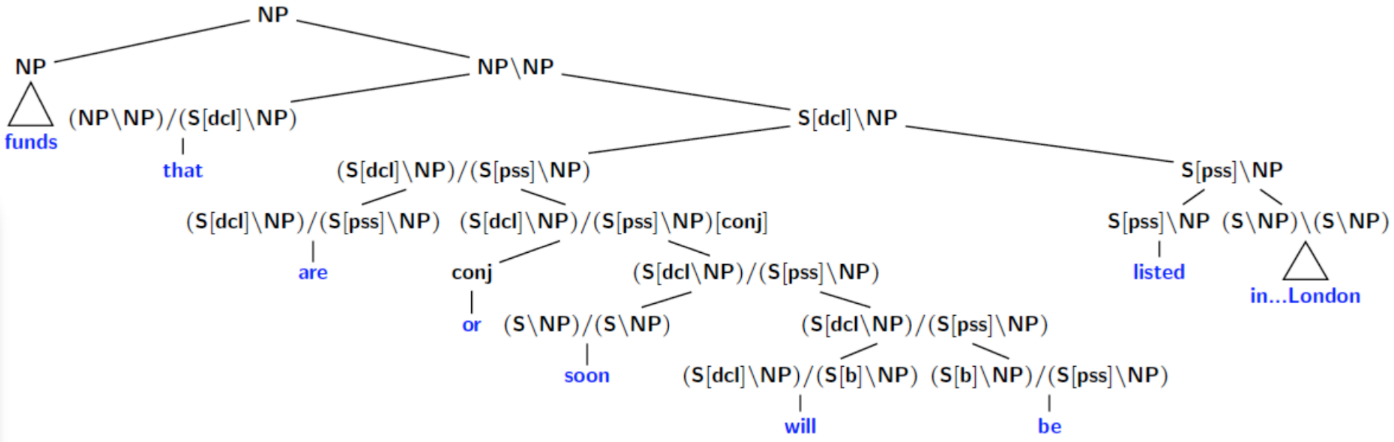
funds that are or soon will be listed in New York or London.



CCGbank
(Hockenmaier and Steedman 2002, 2007; Hockenmaier 2003)

Translation of Penn Treebank to Combinatory Categorial Grammar (CCG)

Enabled wide-coverage CCG parsing and CCG-based semantic analyzers (Boxer)



that	((NP\NP)/(S[dcl]\NP))	funds	are, will	
are	((S[dcl]\NP)/(S[pss]\NP))	*funds*	*listed*	
soon	((S\NP)/(S\NP))		will	
will	((S[dcl]\NP)/(S[b]\NP))	*funds*	be	
be	((S[b]\NP)/(S[pss]\NP))		listed	
listed	(S[pss]\NP)	*funds*		
in	(((S\NP)\(S\NP))/NP)		listed	York, London

What does it take to *'understand'* language?

People are shopping groceries
in a supermarket

No

Yes

Natural language understanding involves the ability to **connect language to the world** (“grounding”)

How do you get a computer to describe images?

How would you describe this image?

A boy in a yellow uniform carrying a football is blocking another boy in a blue uniform.

yes

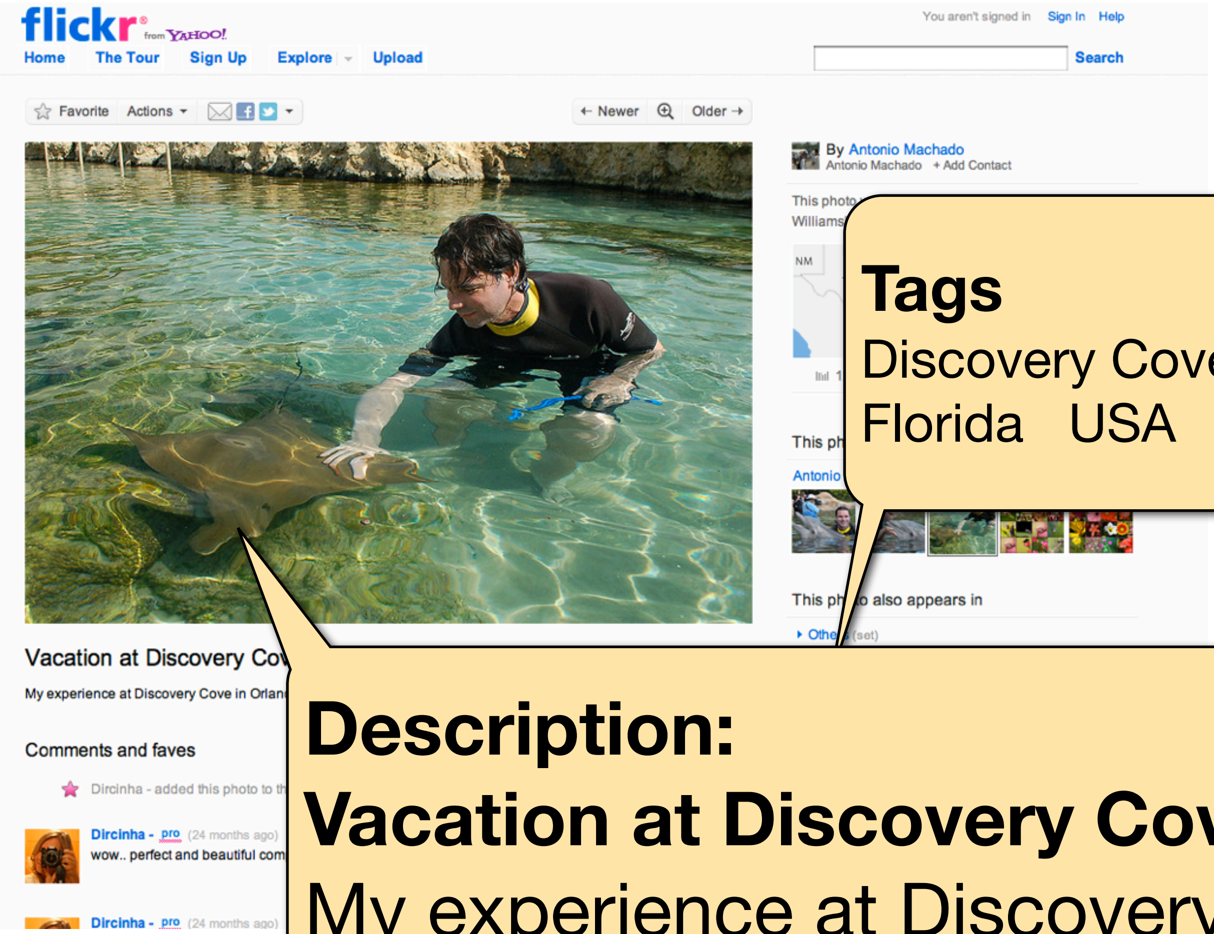
Two boys are playing football

yes

A dog is running on the beach.

no

But people don’t write such captions:



Tags
Discovery Cove Férias Orlando
Florida USA EUA Vacations

Description:
Vacation at Discovery Cove
My experience at Discovery Cove in Orlando, FL

We want generic, conceptual captions

[Shatford, Jaimes et al., Hollink et al.]

Conceptual captions
… describe the depicted **entities, events, scenes**
… only describe **what can be seen in the image**
… may **be more or less specific**

Generic captions **don't refer to named entities**
('a boy', not 'Kevin')

We need to crowdsource captions: Flickr8K/Flickr30K

[Rashtchian et al, 2010, Farhadi et al, 2010, Hodosh et al, 2013]

Four basketball players in action.

Young men playing basketball in a competition.

Four men playing basketball, two from each team.

Two boys in green and white uniforms play basketball with two boys in blue and white uniforms.

A player from the white and green highschool team dribbles down court defended by a player from the other team.

Image description as ranking

Two boys are playing football.

andles.

A little girl is enjoying the swings

A little girl is enjoying the swings

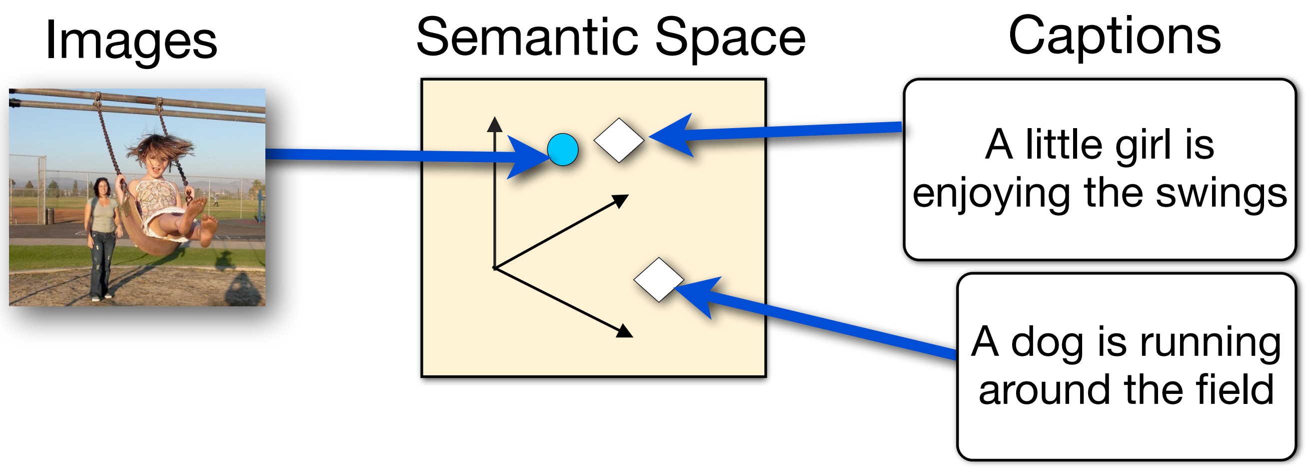
A motorbike is racing around a track.

A boy in a yellow uniform

An elephant is being washed.

Rank a pool of unseen sentences based on how well they describe an (unseen) image.

Ranking by mapping images and sentences to a common semantic vector space



Hodosh, Young, Hockenmaier 2013:
Map images and sentences to a shared vector space,
e.g. by (Kernel) Canonical Correlation Analysis, (K)CCA.
Rank sentences by their distance to the query image.

KCCA results: Image description

A girl wearing a yellow shirt and sunglasses smiles.

A man climbs up a sheer wall of ice.

A child jumping on a tennis court.

Basketball players in action.

Hodosh, Young, Hockenmaier 2013
No object/scene detectors,
No neural nets/deep learning,
just pyramid kernels over low-level visual features
(SIFT, texture, color)

So, we can go home, right?

Do these models actually understand language (or images)?

Hodosh & Hockenmaier 2016

Binary Forced-Choice Tasks

GOLD

A. There is a woman riding a bike down the road and she popped a wheelie.

DISTRACTOR

B. Two men in jeans and jackets are walking down a small road.

How often does an **off-the-shelf system** score the original gold caption higher than a distractor caption?

In each task, **gold and distractor differ systematically**

Models pick up on scene terms,
but don't understand who does what to whom
(Hodosh & Hockenmaier 2016)

What does this mean?

Learning to associate images with simple sentences that describe them is clearly a **much easier task** than we thought not too long ago.

But image captioning systems (from ~2015) didn't actually 'understand' how to associate simple sentences with images

The ELIZA effect is well and alive...

ELIZA effect (Weizenbaum 1966)

It's easy to overestimate
the abilities of NLP systems

Phrase Grounding as a harder challenge

Which child is being pushed?

Phrase grounding may require more sophisticated language and image understanding.

A woman pushes a child on a swing while another child looks on.

Flickr30K Entities

[Plummer, Wang, Cervantes, Caicedo, Hockenmaier, Lazebnik, 2015]

Flickr30k Entities augments Flickr30k with **267,000 bounding boxes** and **244,000 coreference chains** for all mentioned entities.
Annotation was done via crowdsourcing.

A man with pierced ears is wearing glasses and an orange hat.
A man with glasses is wearing a beer can crocheted 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.

What does it take to *'understand'* language?

People are shopping groceries in a supermarket

They are sitting at desks.
They are walking on the street.
They are buying clothes.
They are at home.

No

They are standing or walking.
They are pushing shopping carts.
They are in an indoor space.
There are aisles of shelves

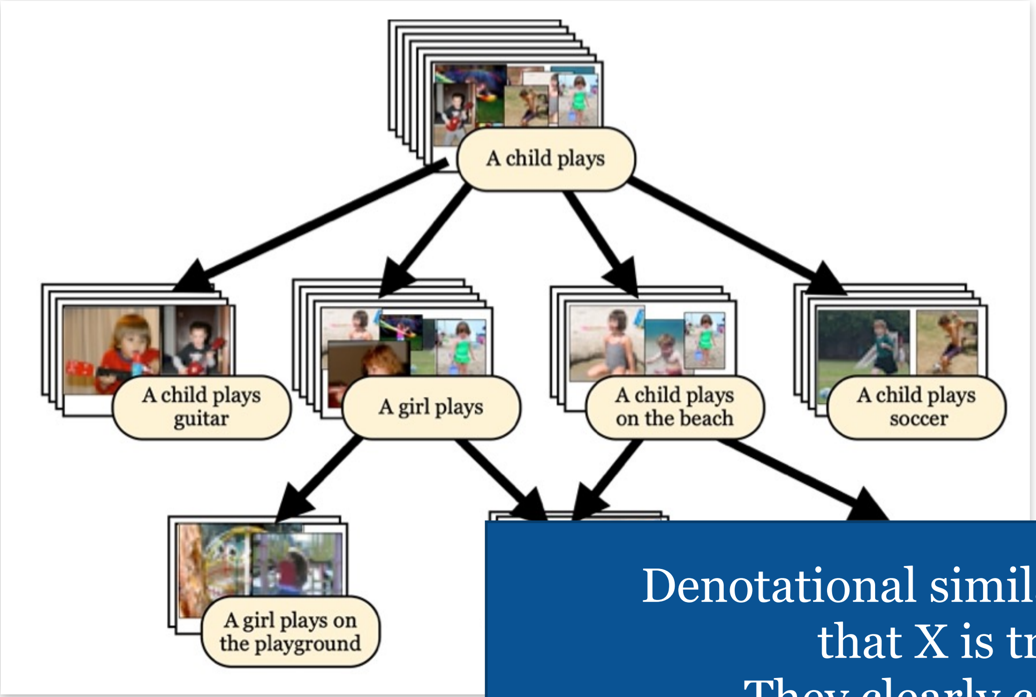
Yes

Natural language understanding involves the ability to

draw inferences

(which may require commonsense/world knowledge)

Flickr30K also allowed us to compute a ”denotation graph” and “denotational similarities”

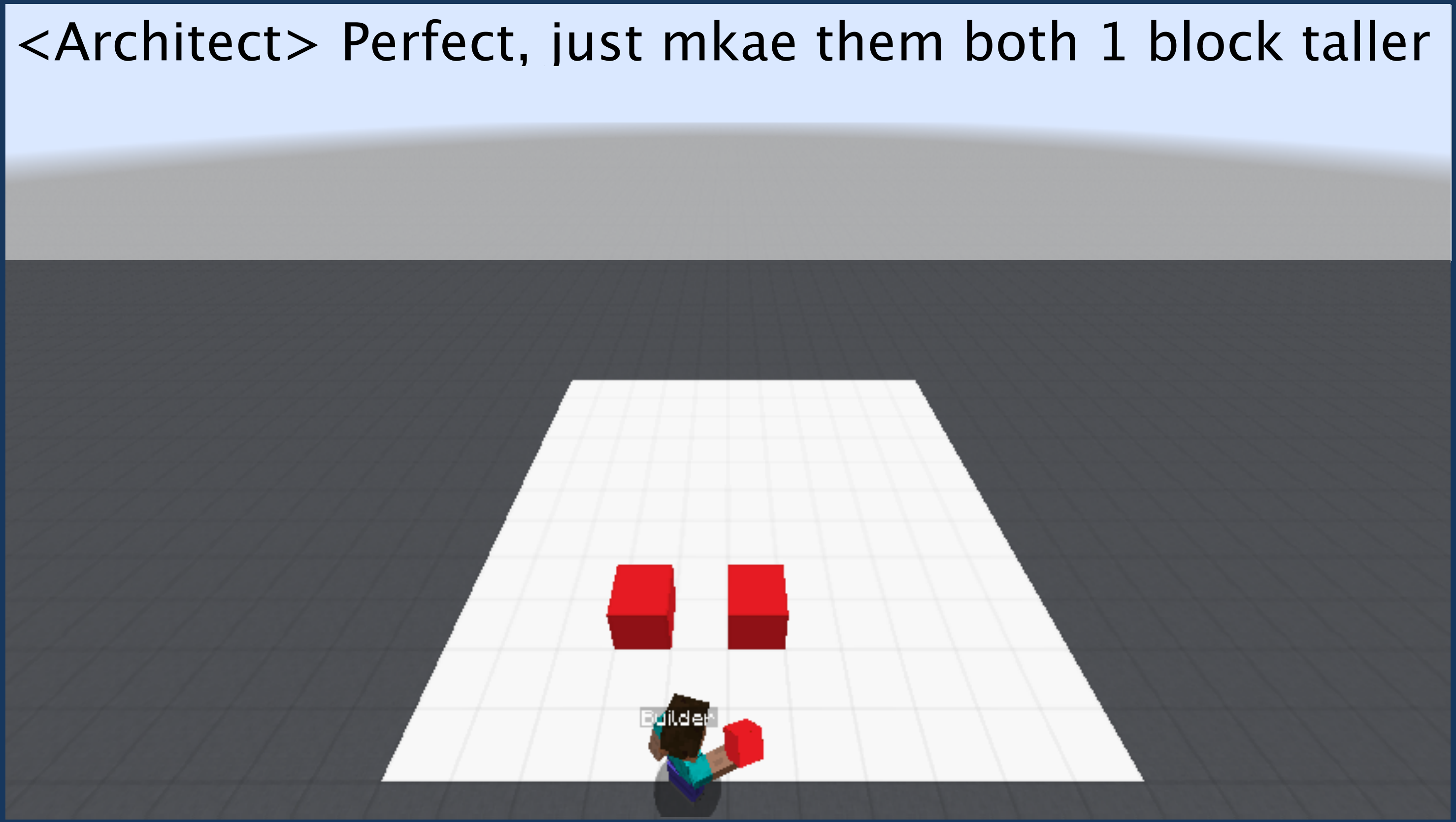


p(VP_1		VP_2)	
p(talk	| engage in conversation)	=	0.79
p(play tennis	| swing racket)	=	0.82
p(stand	| wait for subway)	=	0.58
p(sit	| ride subway)	=	0.56
p(stand	| lean against building)	=	0.53
p(shave	| look in mirror)	=	0.41
p(dig hole	| use shovel)	=	0.38
			0.38

Denotational similarities, e.g. P(X | Y), express the probability that X is true if Y is true in the same situation.
They clearly capture some commonsense knowledge, and can be useful for textual entailment.

However, 30K images with 5 captions is not nearly enough data

What does it take to "understand" language?



Natural language understanding
being able to collaborate with others
(e.g. to give and follow instructions)

How do you get a computer to give or follow instructions?

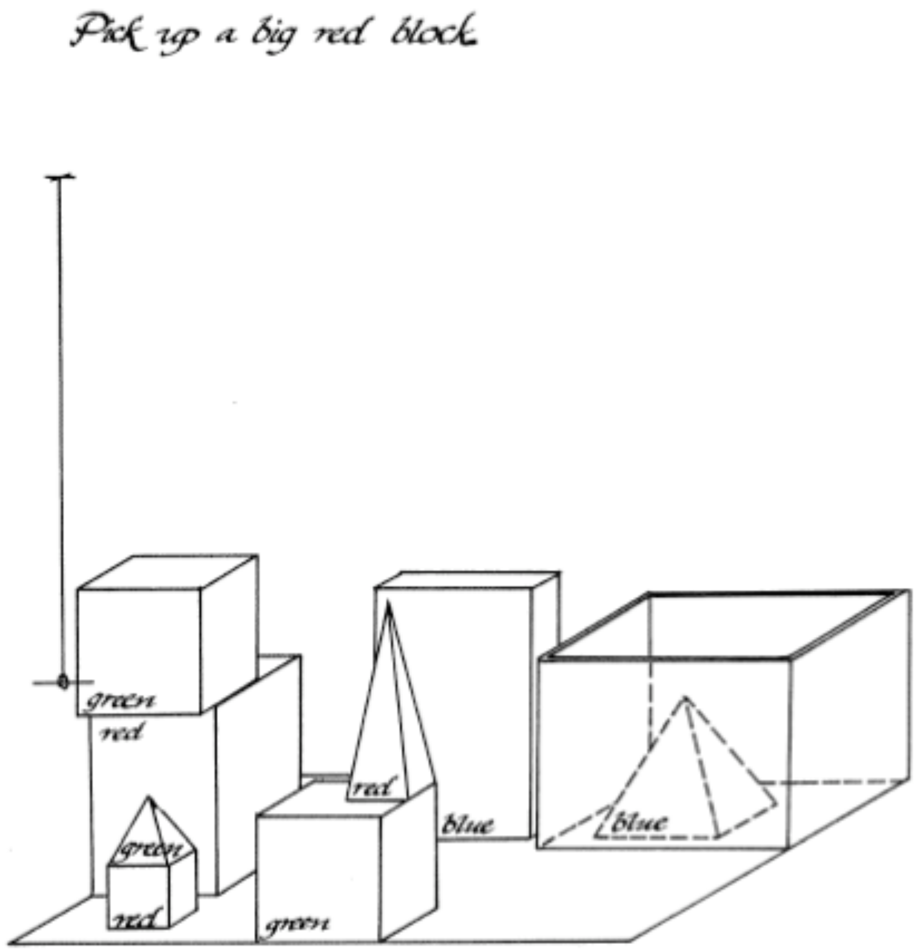
Communication and Collaborative Construction in Minecraft and other BlocksWorlds

Anjali Narayan-Chen, Prashant Jayannavar, Harsha Kokel, Mayukh Das,
Rakib Islam, Julia Bonn, Jon Cai, Susan Brown, Soham Dan,
Jana Doppa, Sriraam Natarajan, Martha Palmer, Dan Roth, Julia Hockenmaier

DARPA's Communicating with Computers (CwC) program

CwC aims to enable
symmetric **communication between computers and people**
in collaborative contexts.

The Blocks World use case:
Humans and machines communicate
to build a given target structure
with toy blocks.



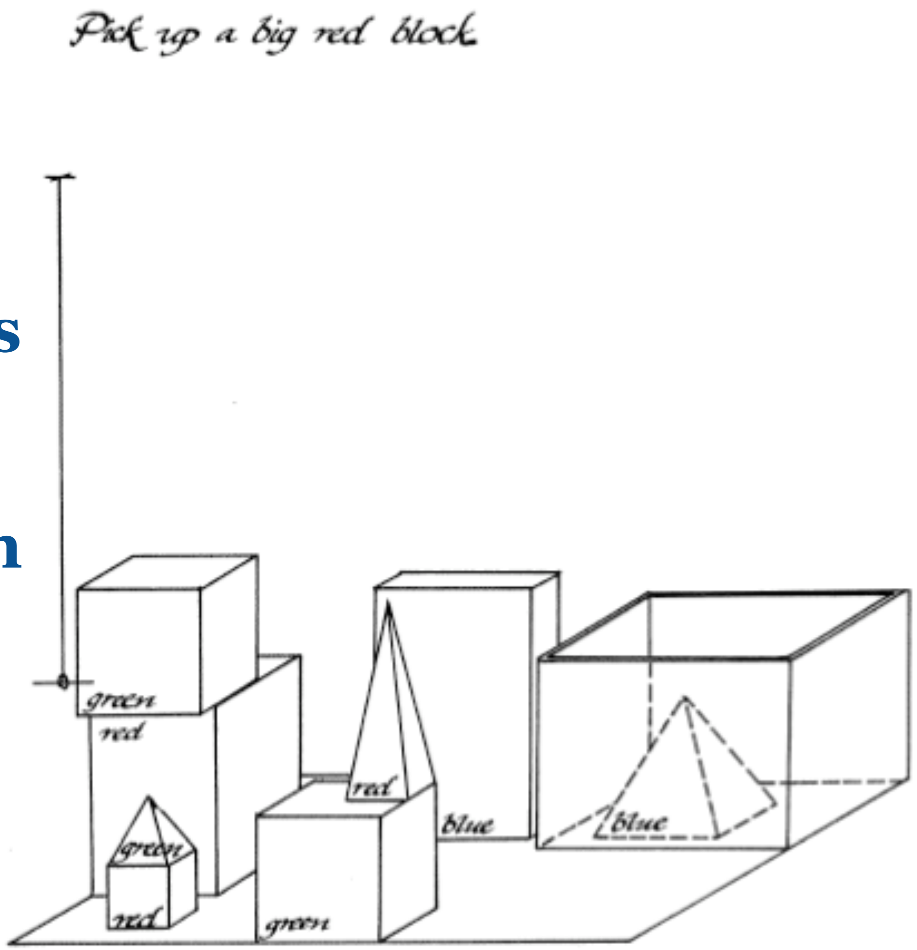
Blocks World: Winograd's SHRDLU (1971)

SHRDLU had a symbolic representation of a scene with several different types of blocks (to simulate an **immobile robot with an arm**)

Users could **instruct SHRDLU to move blocks** in this scene (and ask questions about the scene)

But SHRDLU was based entirely on **handwritten symbolic rules and domain knowledge**.

Can modern systems learn to perform this task without handwritten rules?



Minecraft as a virtual platform for NLP

Popular multi-player gaming platform where **avatars navigate in a 3D world** and **manipulate block-like materials**

Microsoft's **Project Malmo API** makes it possible to use Minecraft for reinforcement learning and other AI research.

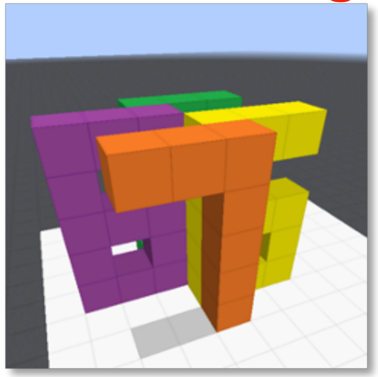
We show that this makes Minecraft a great virtual platform to study **interactive, situated language generation & understanding**.

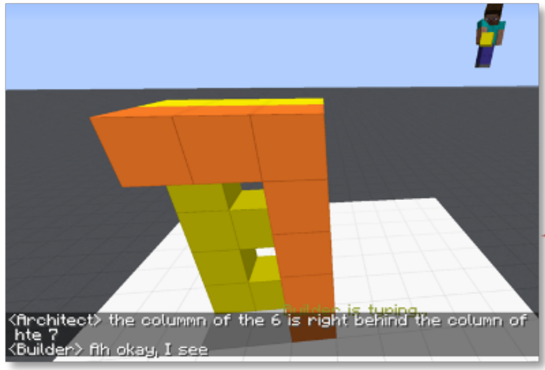
We can use Minecraft to simulate a **Blocks World for embodied agents**

The Minecraft Collaborative Building Task

The Architect

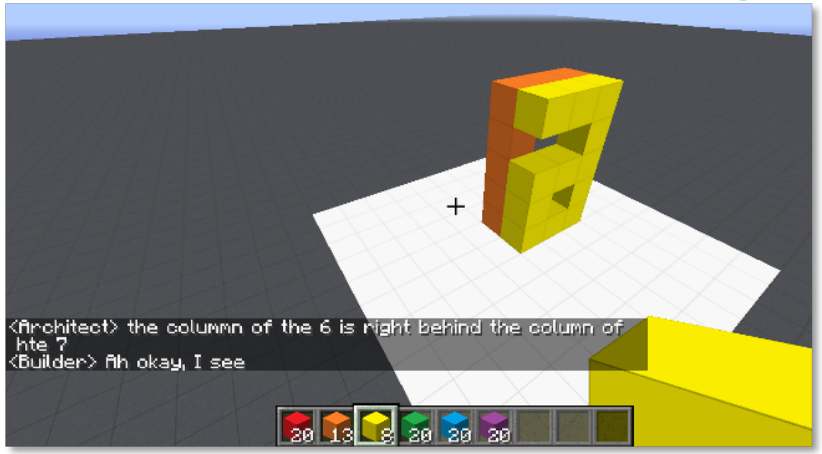
knows the Target **observes the Builder**





The Builder

has to build a copy of the Target



Chat Interface

A: In about the middle build a column five tall

A: then two more to the left of the top to make a 7

A: now a yellow 6

A: the long edge of the 6 aligns with the stem of the 7 and faces right

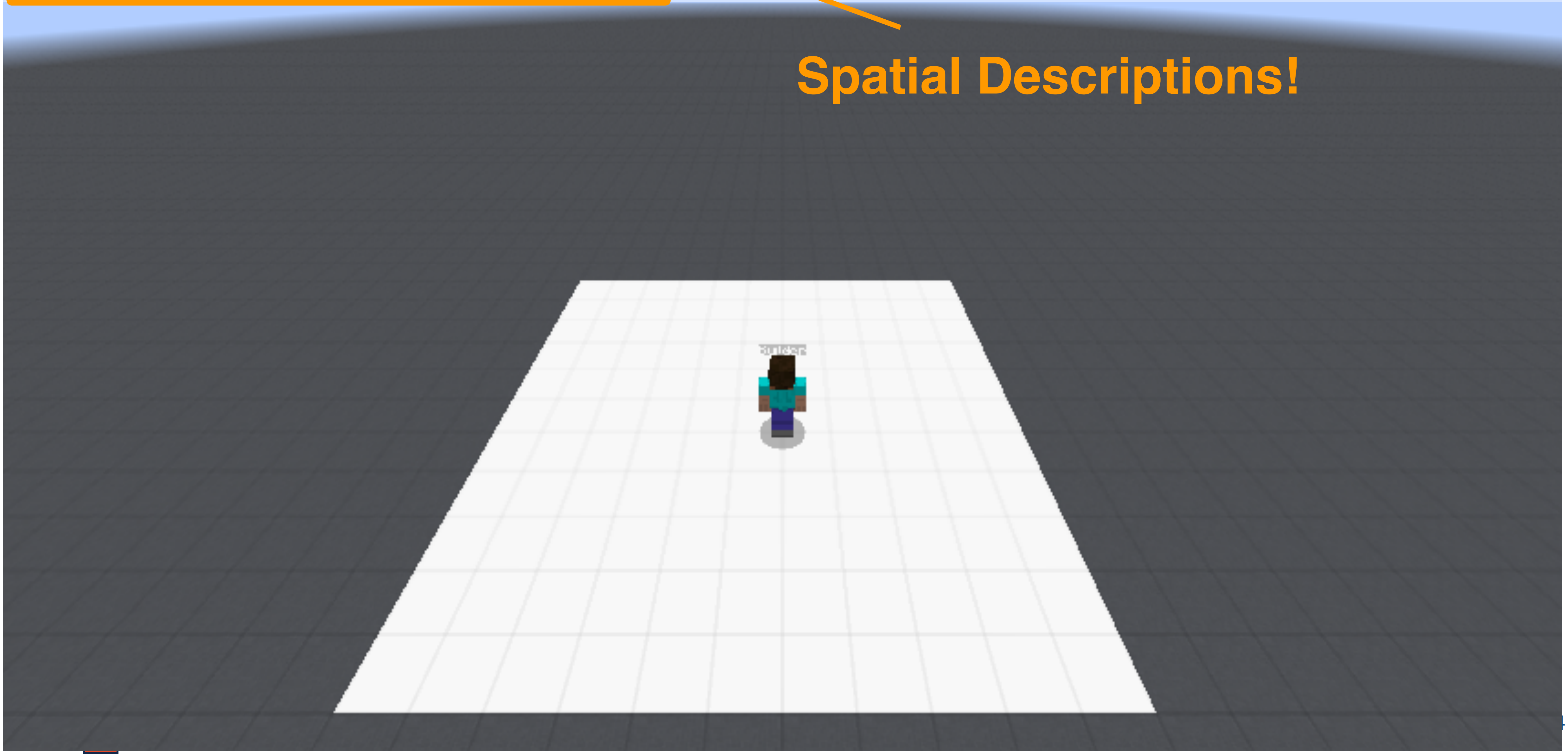
B: where does the 6 start?

A: behind the 7 from your perspective

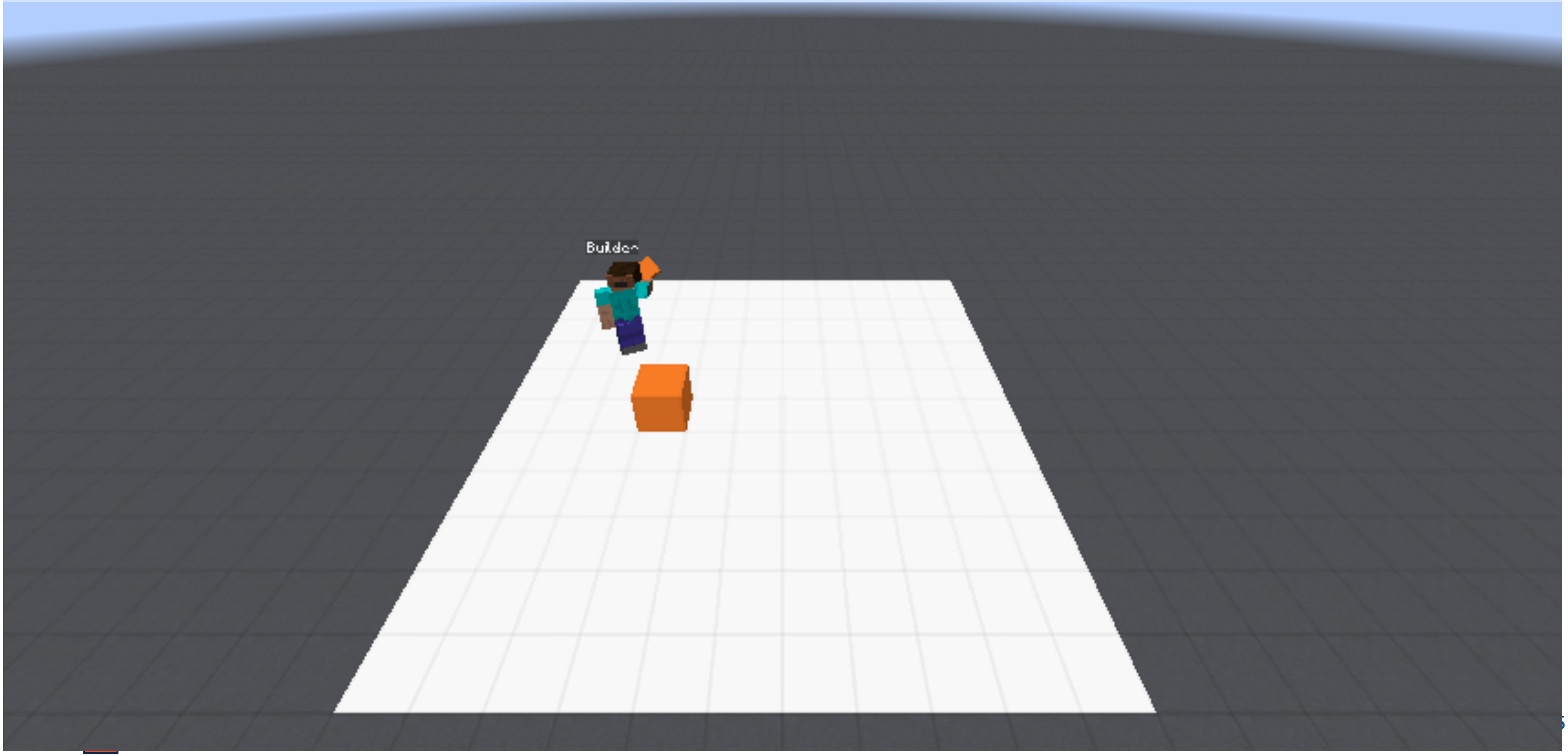
How do People Perform This Task?

<Architect> go the middle and place an orange block two spaces to the left

Spatial Descriptions!



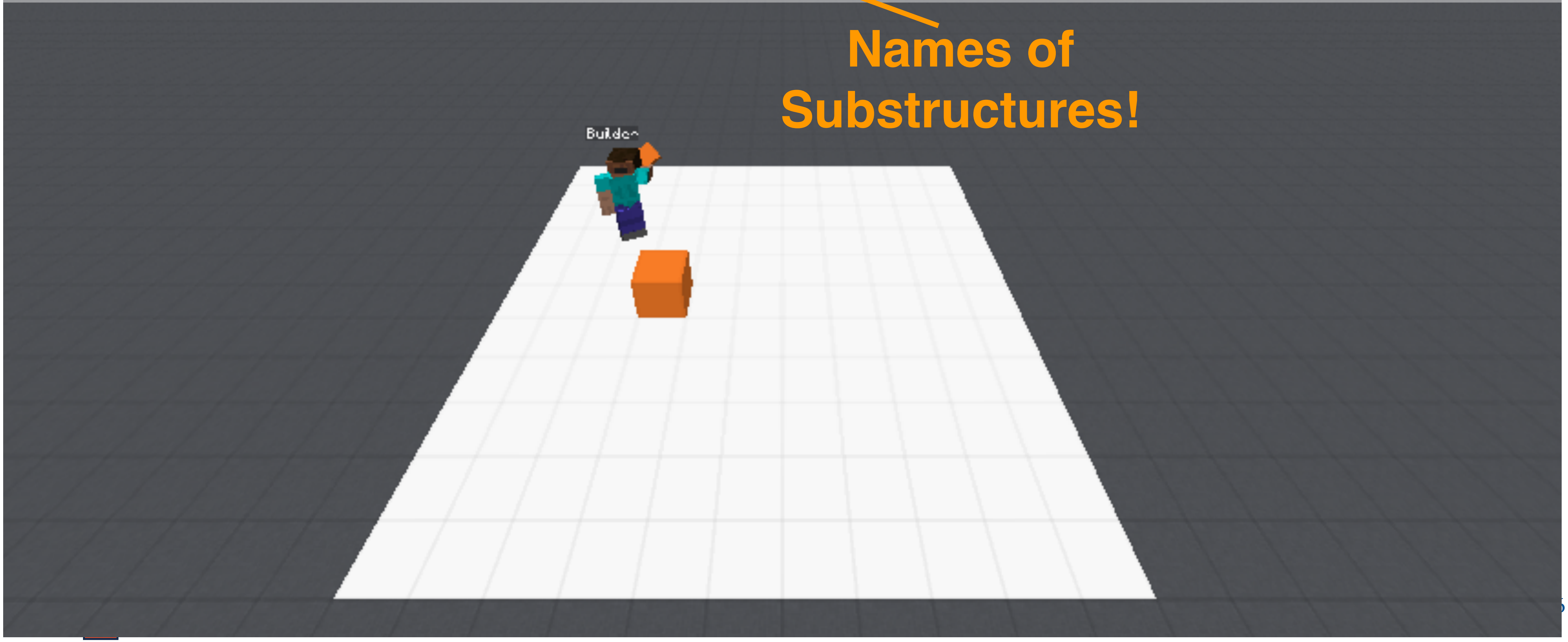
<Architect> go the middle and place an orange block two spaces to the left

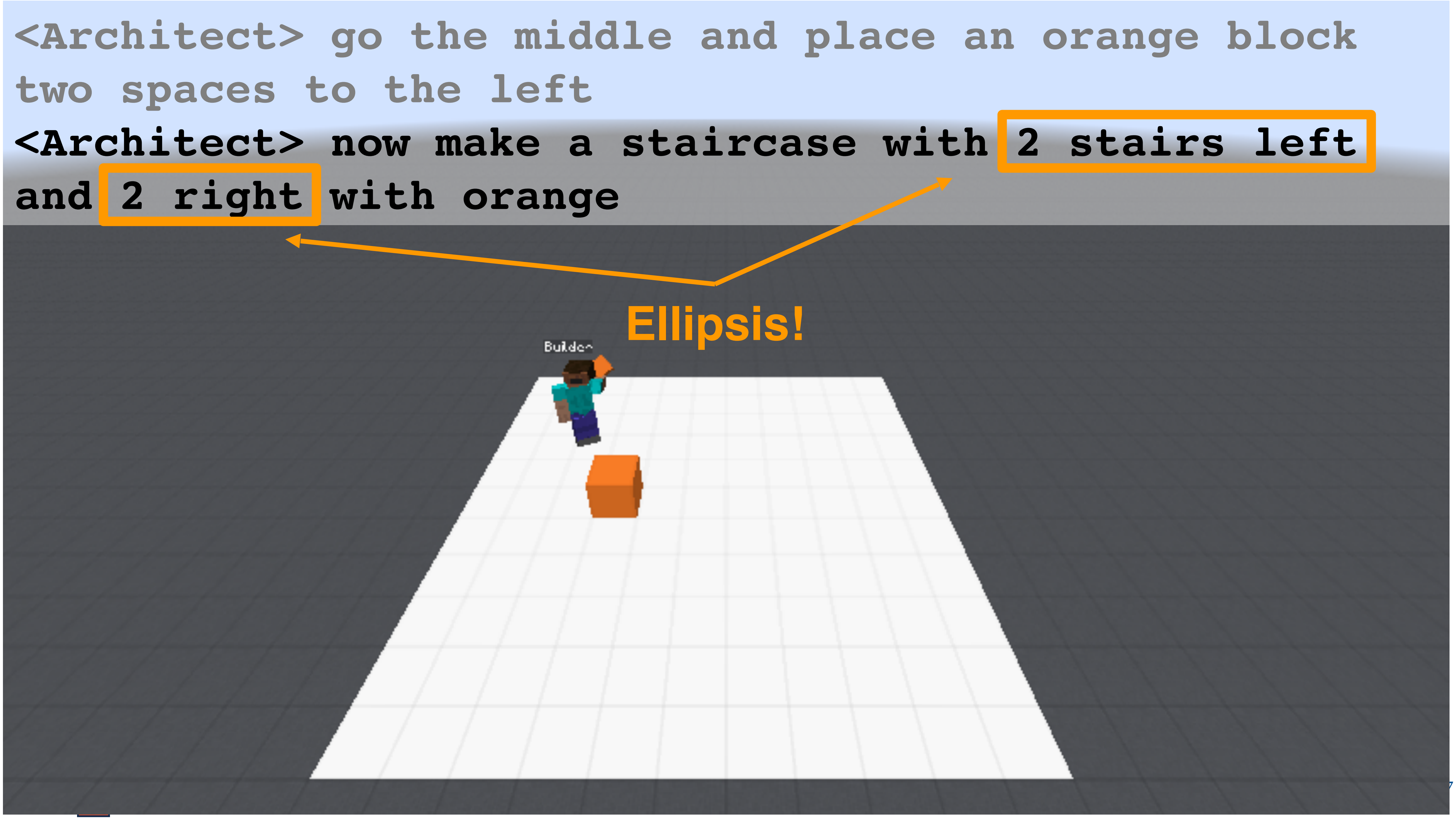


<Architect> go the middle and place an orange block two spaces to the left

<Architect> now make a staircase with 2 stairs left and 2 right with orange

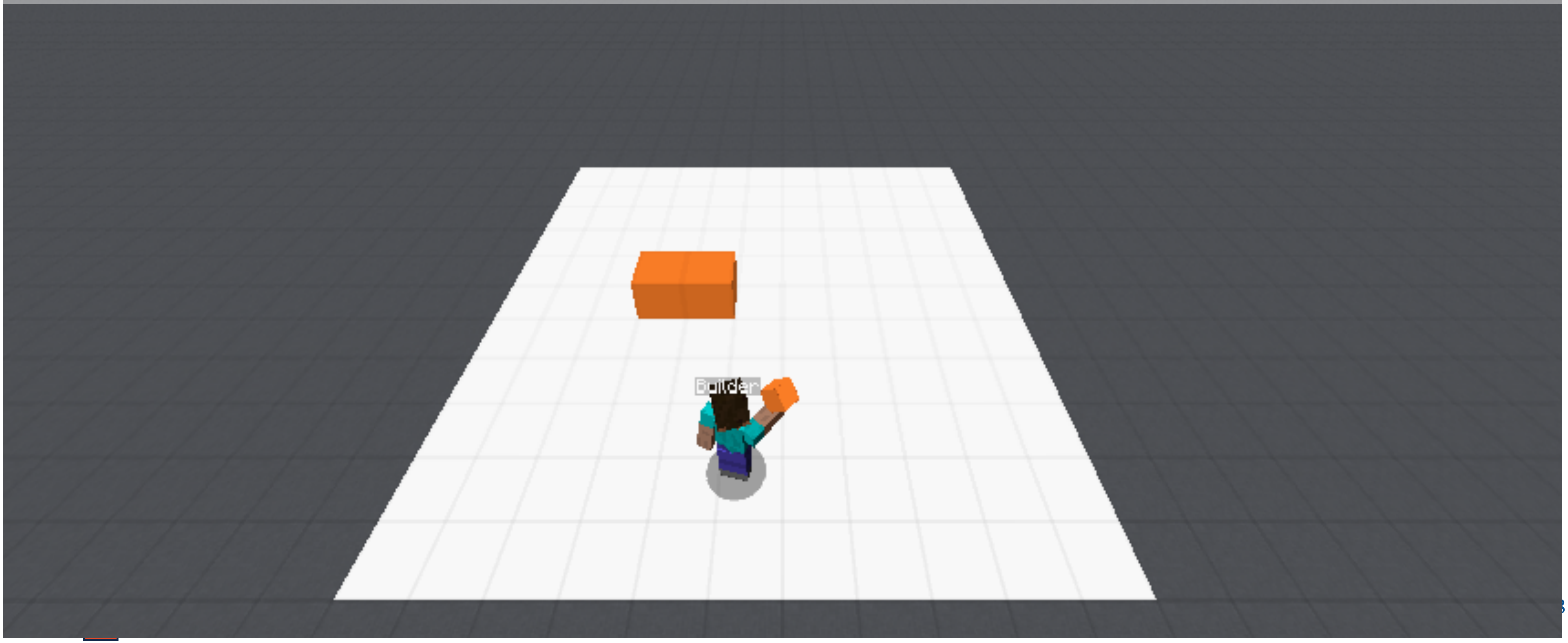
Names of Substructures!

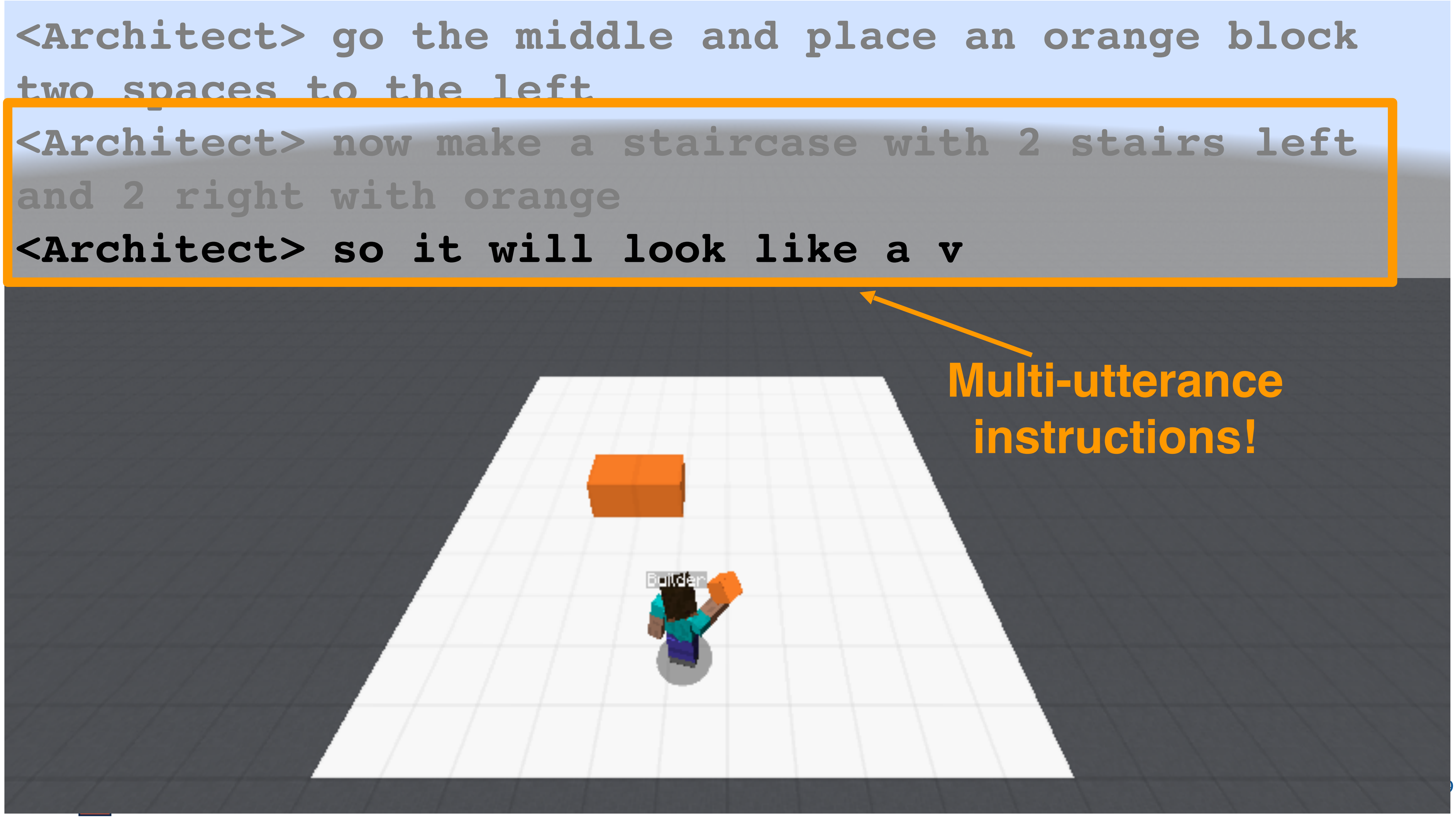


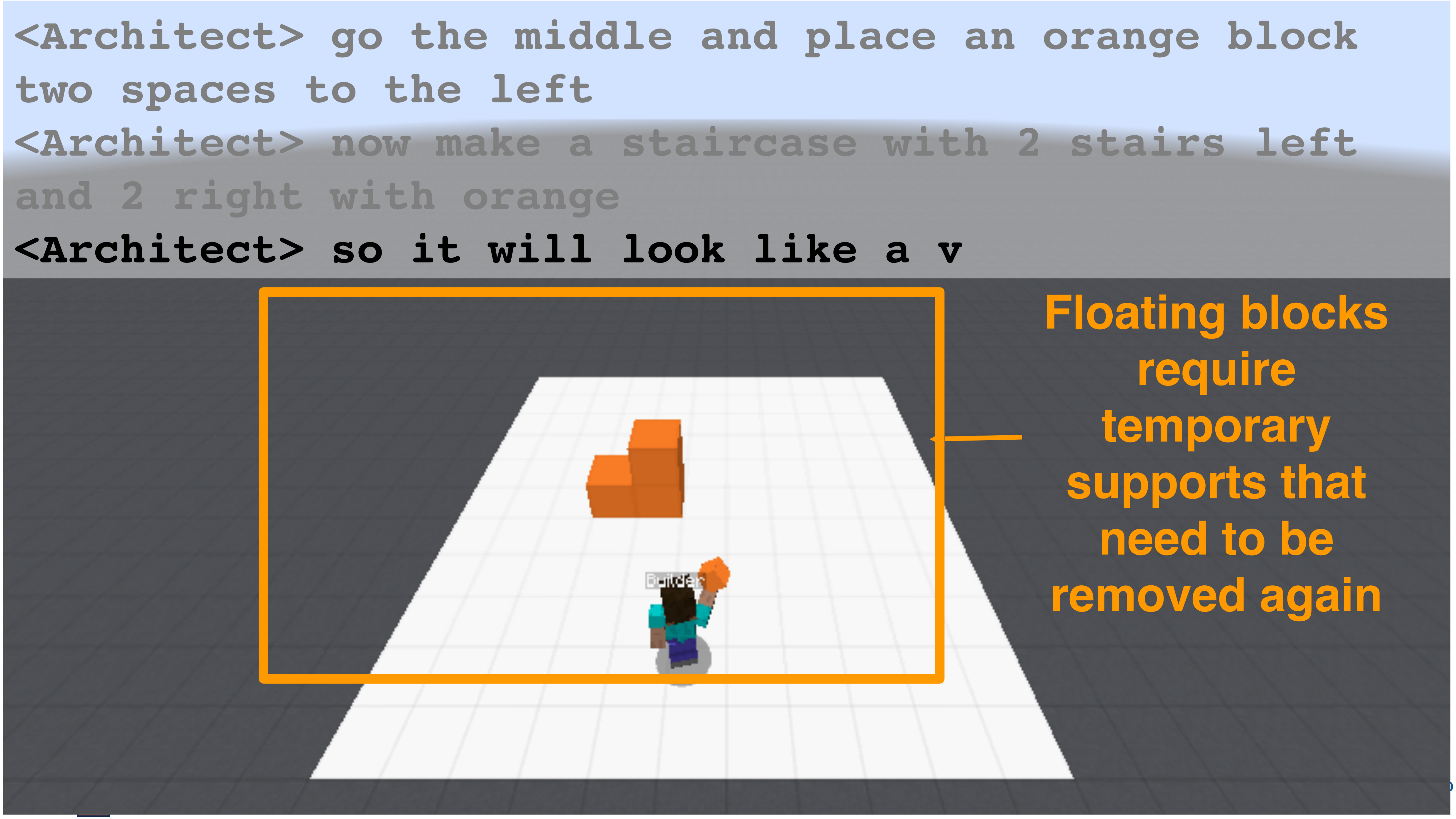

<Architect> go the middle and place an orange block two spaces to the left
<Architect> now make a staircase with 2 stairs left and 2 right with orange
Ellipsis!
Builder

<Architect> go the middle and place an orange block two spaces to the left

<Architect> now make a staircase with 2 stairs left and 2 right with orange

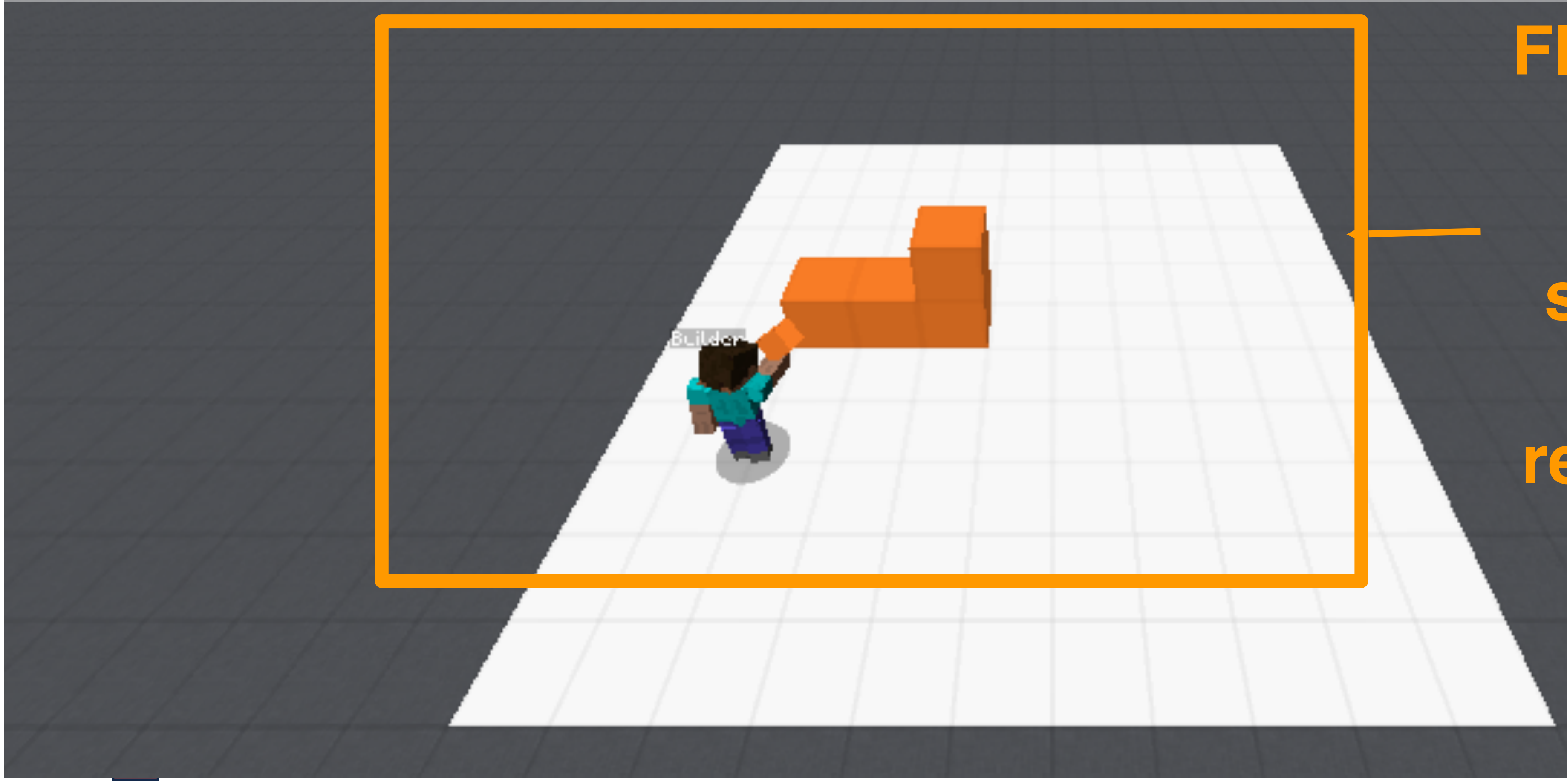



<Architect> go the middle and place an orange block two spaces to the left
<Architect> now make a staircase with 2 stairs left and 2 right with orange
<Architect> so it will look like a v
Builder
Multi-utterance instructions!

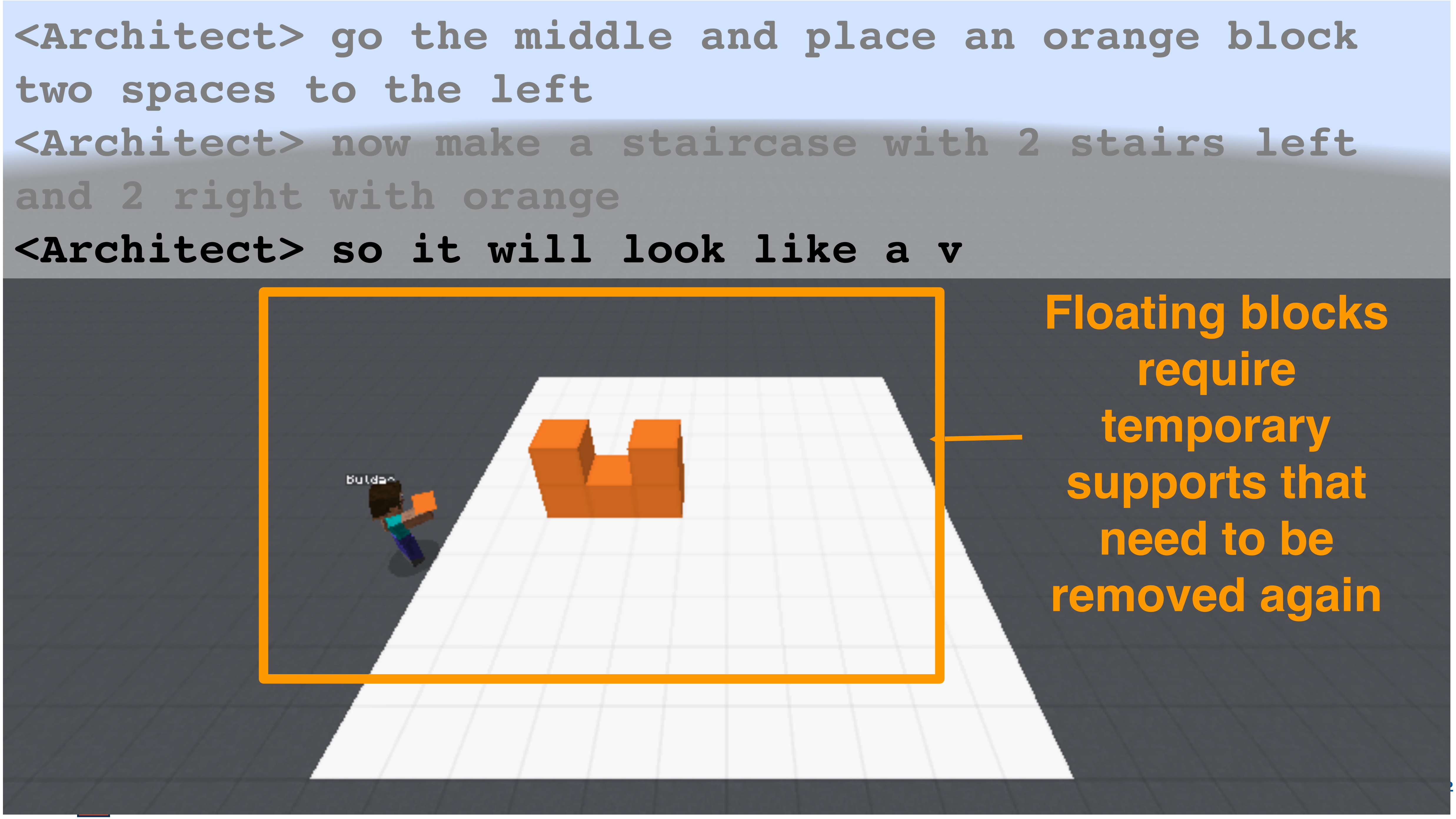


<Architect> go the middle and place an orange block two spaces to the left
<Architect> now make a staircase with 2 stairs left and 2 right with orange
<Architect> so it will look like a v
Builder
Floating blocks require temporary supports that need to be removed again

<Architect> go the middle and place an orange block two spaces to the left
<Architect> now make a staircase with 2 stairs left and 2 right with orange
<Architect> so it will look like a v



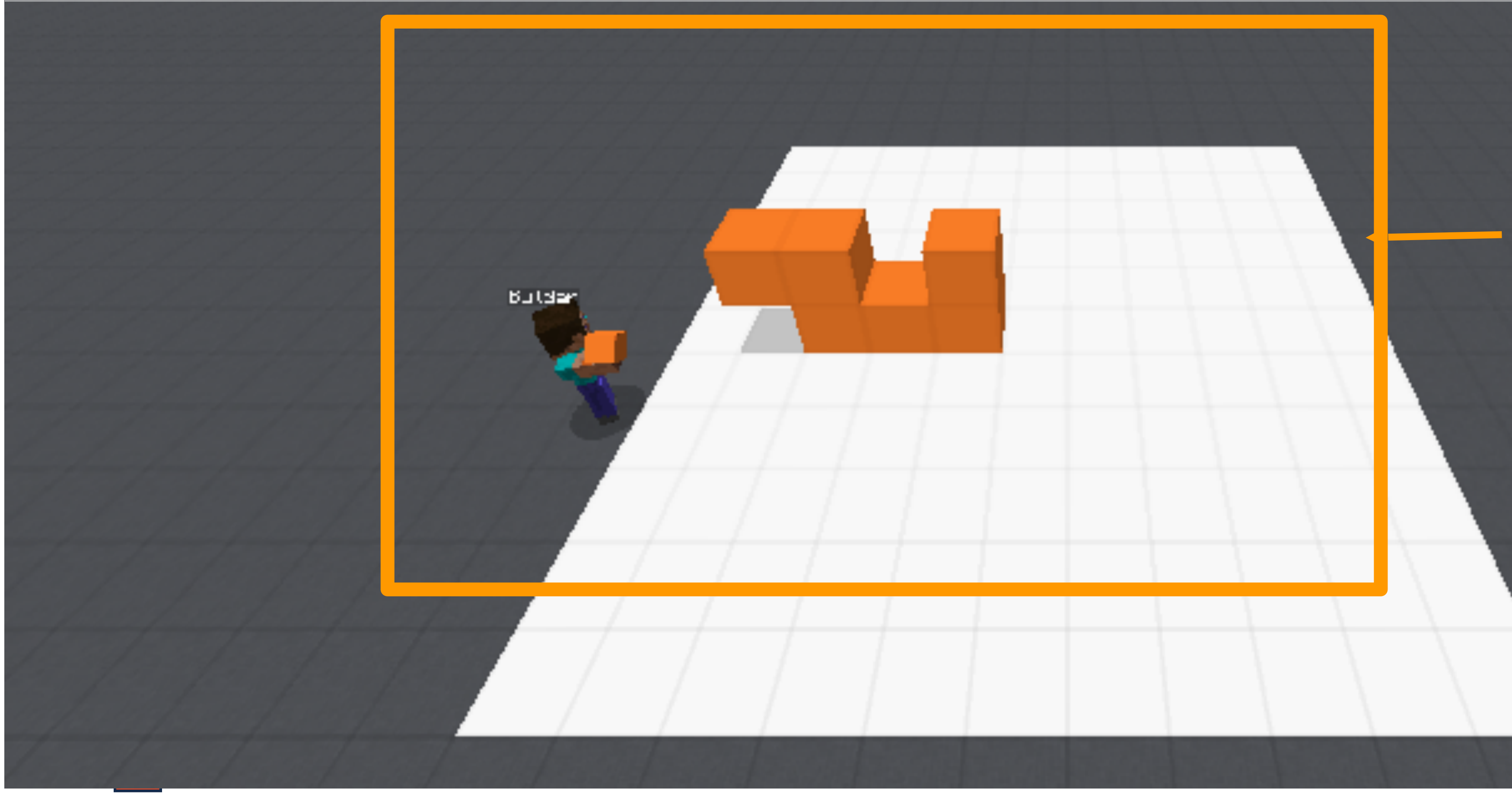
Floating blocks require temporary supports that need to be removed again


<Architect> go the middle and place an orange block two spaces to the left
<Architect> now make a staircase with 2 stairs left and 2 right with orange
<Architect> so it will look like a v
Floating blocks require temporary supports that need to be removed again

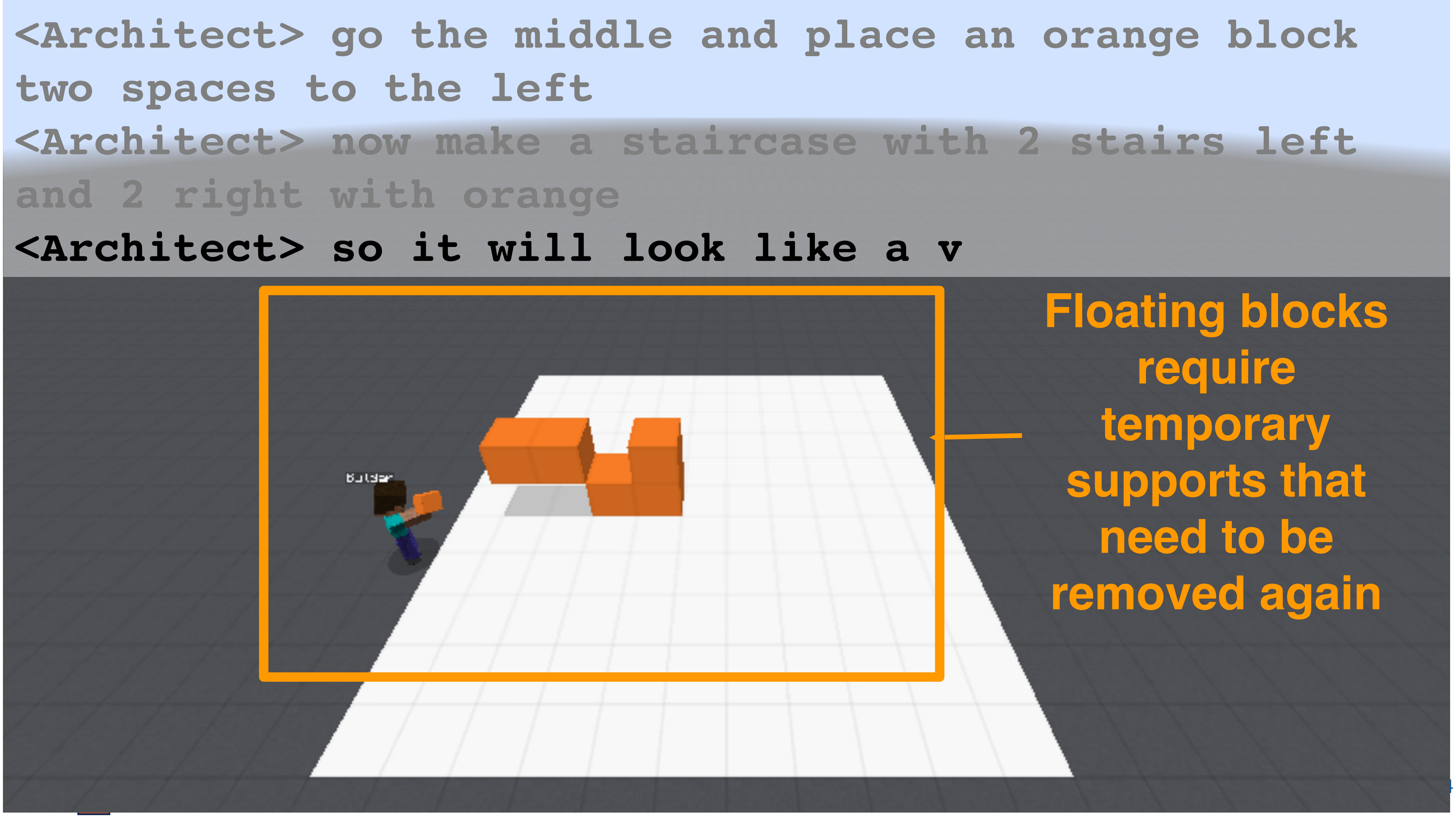
<Architect> go the middle and place an orange block two spaces to the left

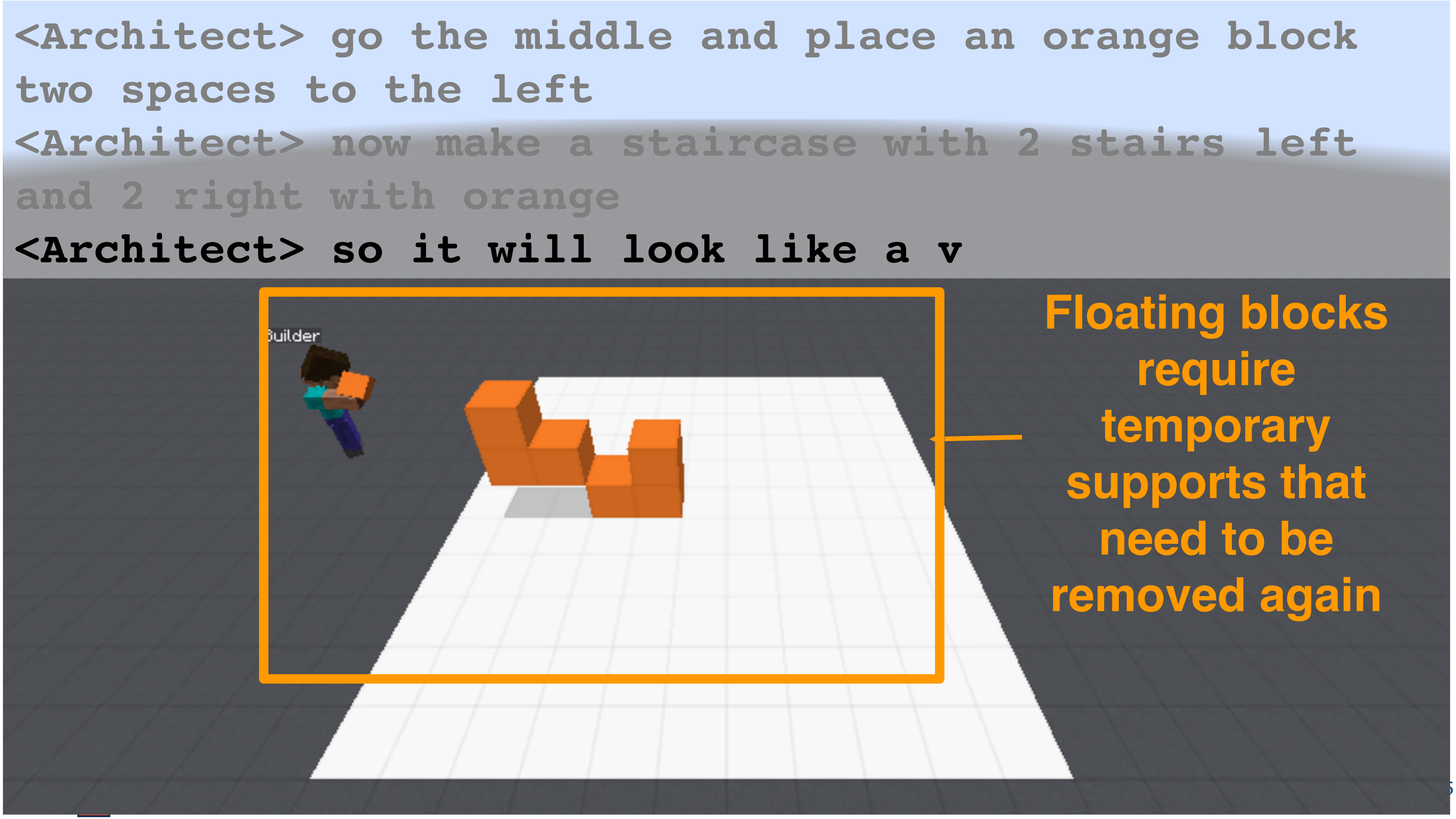
<Architect> now make a staircase with 2 stairs left and 2 right with orange

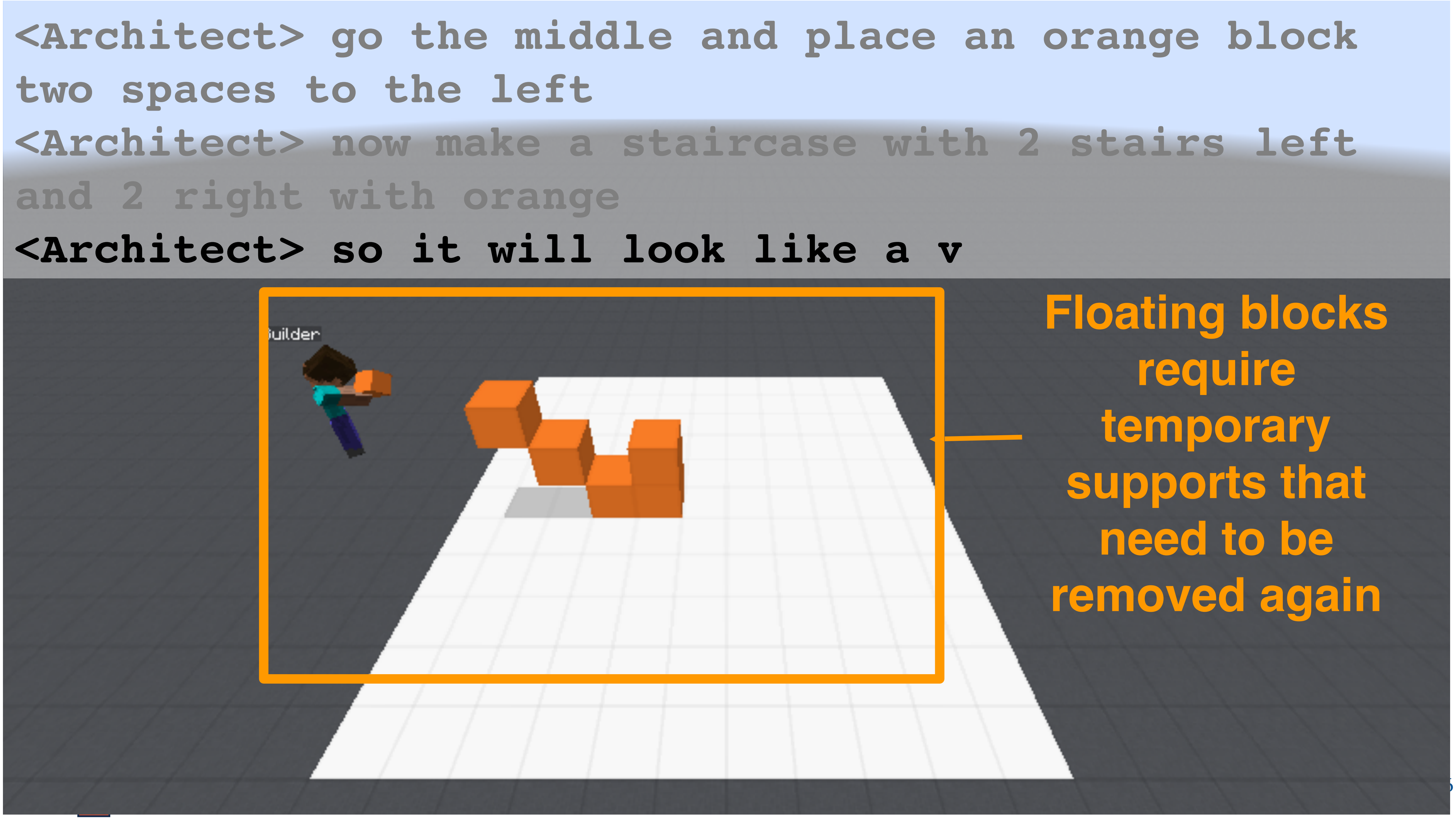
<Architect> so it will look like a v

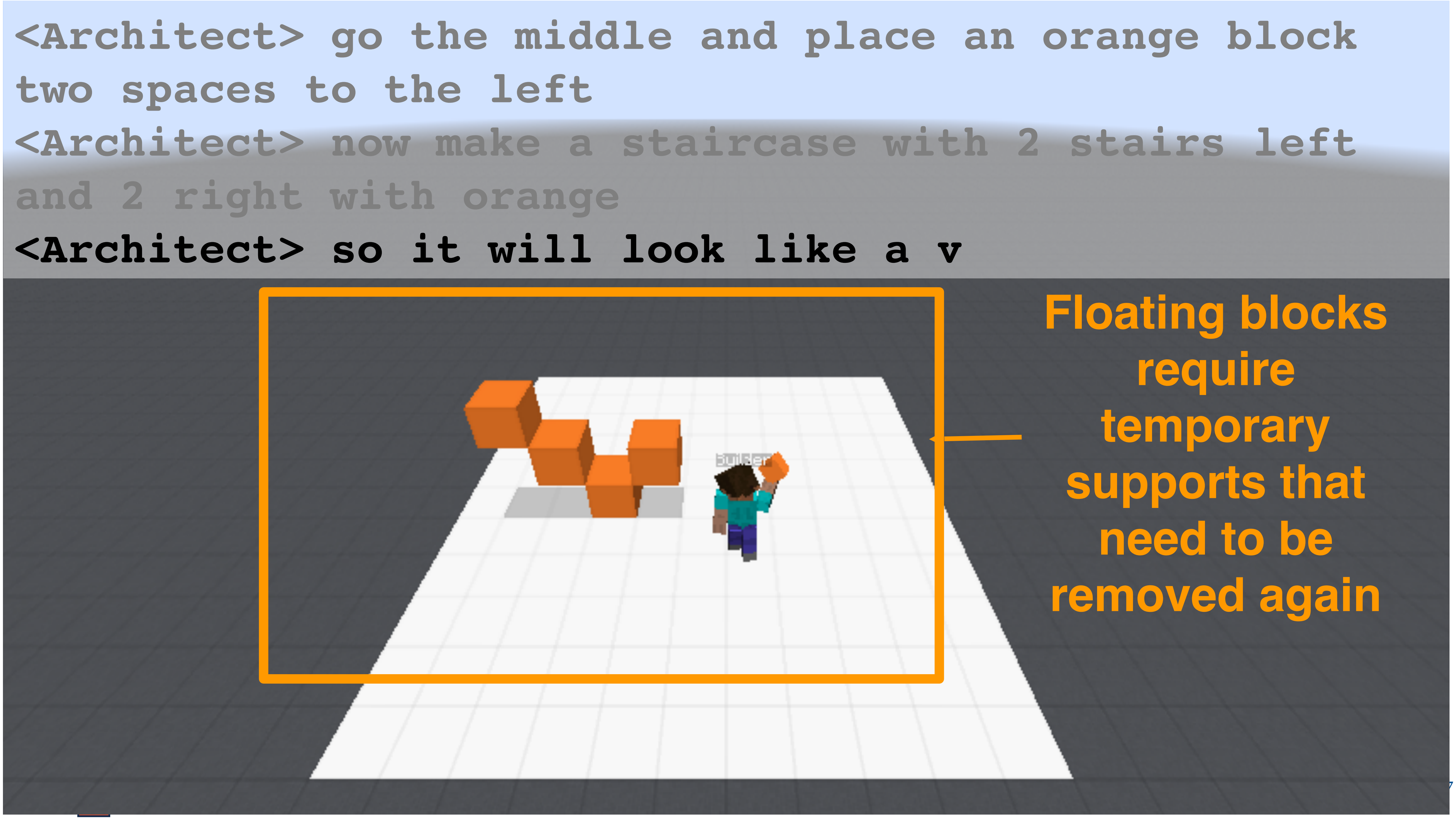


Floating blocks require temporary supports that need to be removed again


<Architect> go the middle and place an orange block two spaces to the left
<Architect> now make a staircase with 2 stairs left and 2 right with orange
<Architect> so it will look like a v
Floating blocks require temporary supports that need to be removed again


<Architect> go the middle and place an orange block two spaces to the left
<Architect> now make a staircase with 2 stairs left and 2 right with orange
<Architect> so it will look like a v
Builder
Floating blocks require temporary supports that need to be removed again

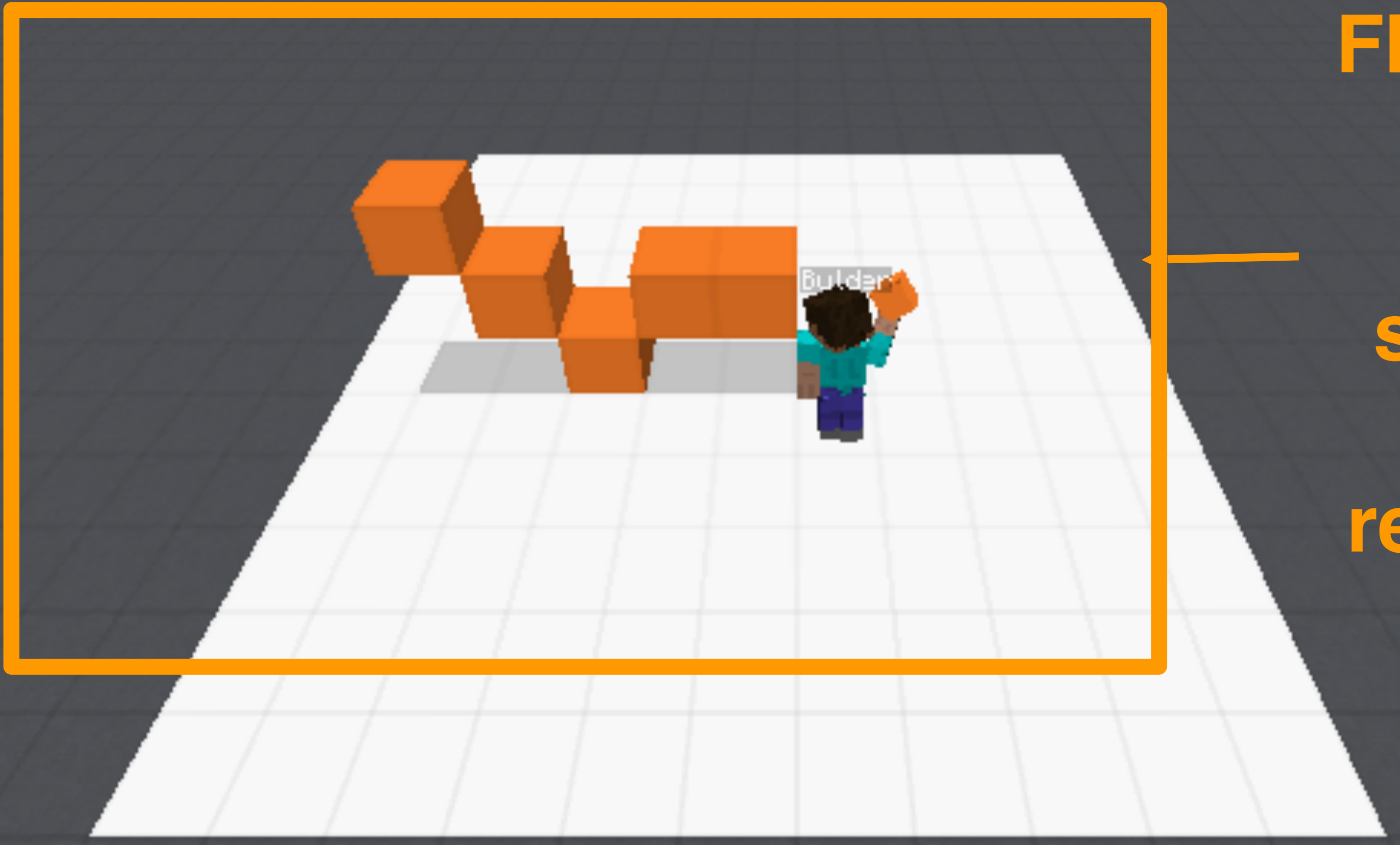

<Architect> go the middle and place an orange block two spaces to the left
<Architect> now make a staircase with 2 stairs left and 2 right with orange
<Architect> so it will look like a v
Builder
Floating blocks require temporary supports that need to be removed again


<Architect> go the middle and place an orange block two spaces to the left
<Architect> now make a staircase with 2 stairs left and 2 right with orange
<Architect> so it will look like a v
Builder
Floating blocks require temporary supports that need to be removed again

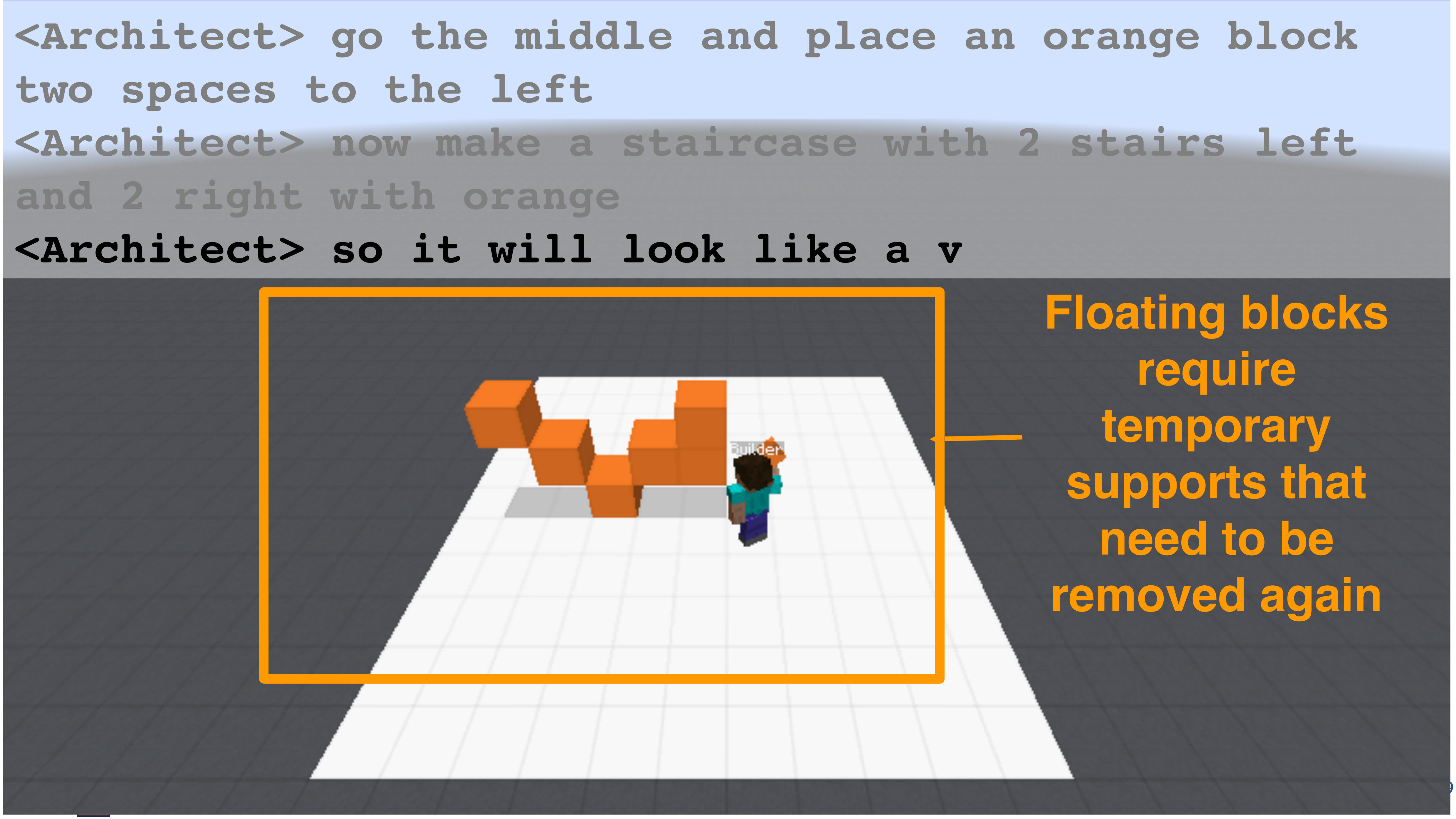
<Architect> go the middle and place an orange block two spaces to the left

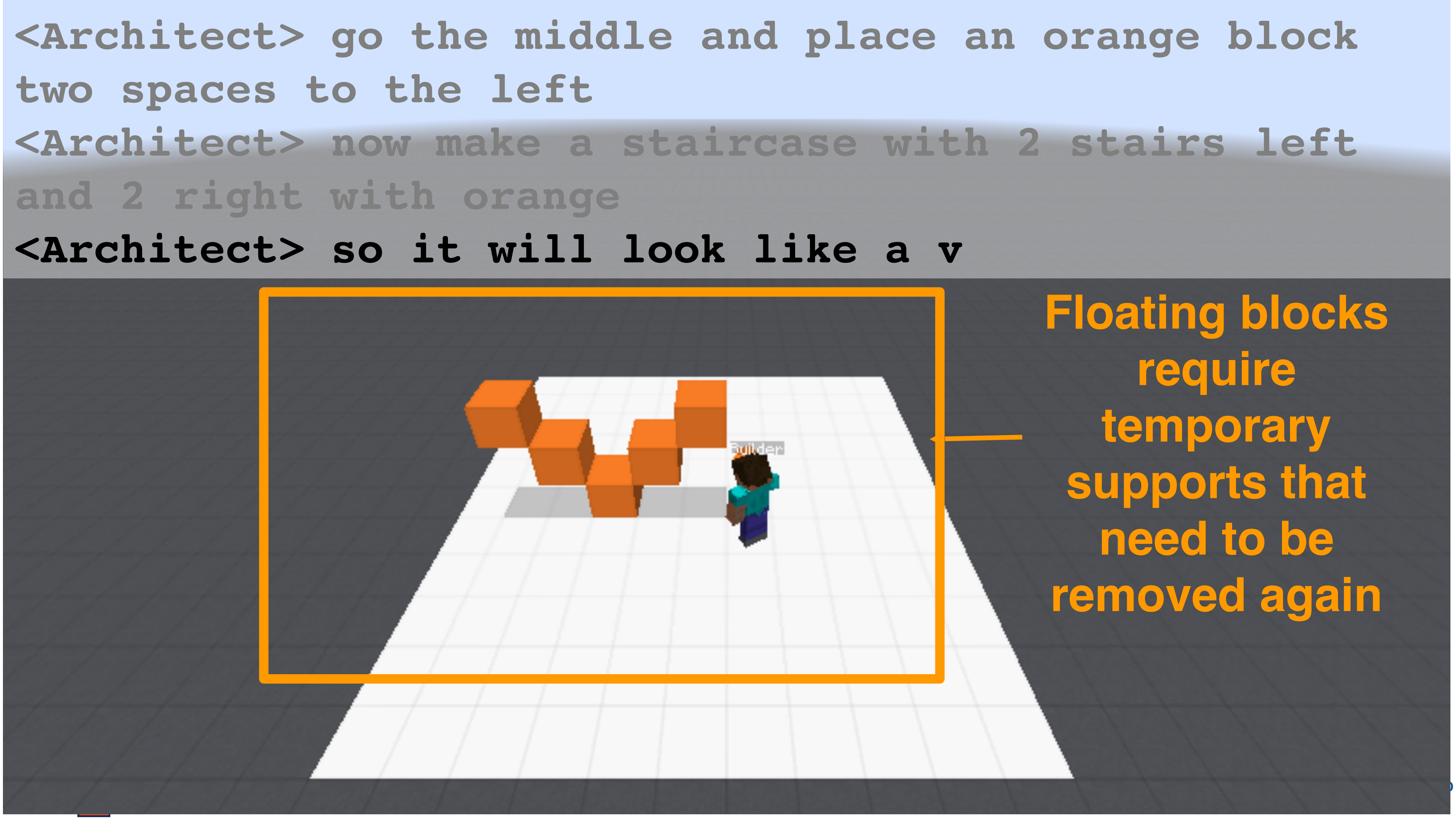
<Architect> now make a staircase with 2 stairs left and 2 right with orange

<Architect> so it will look like a v

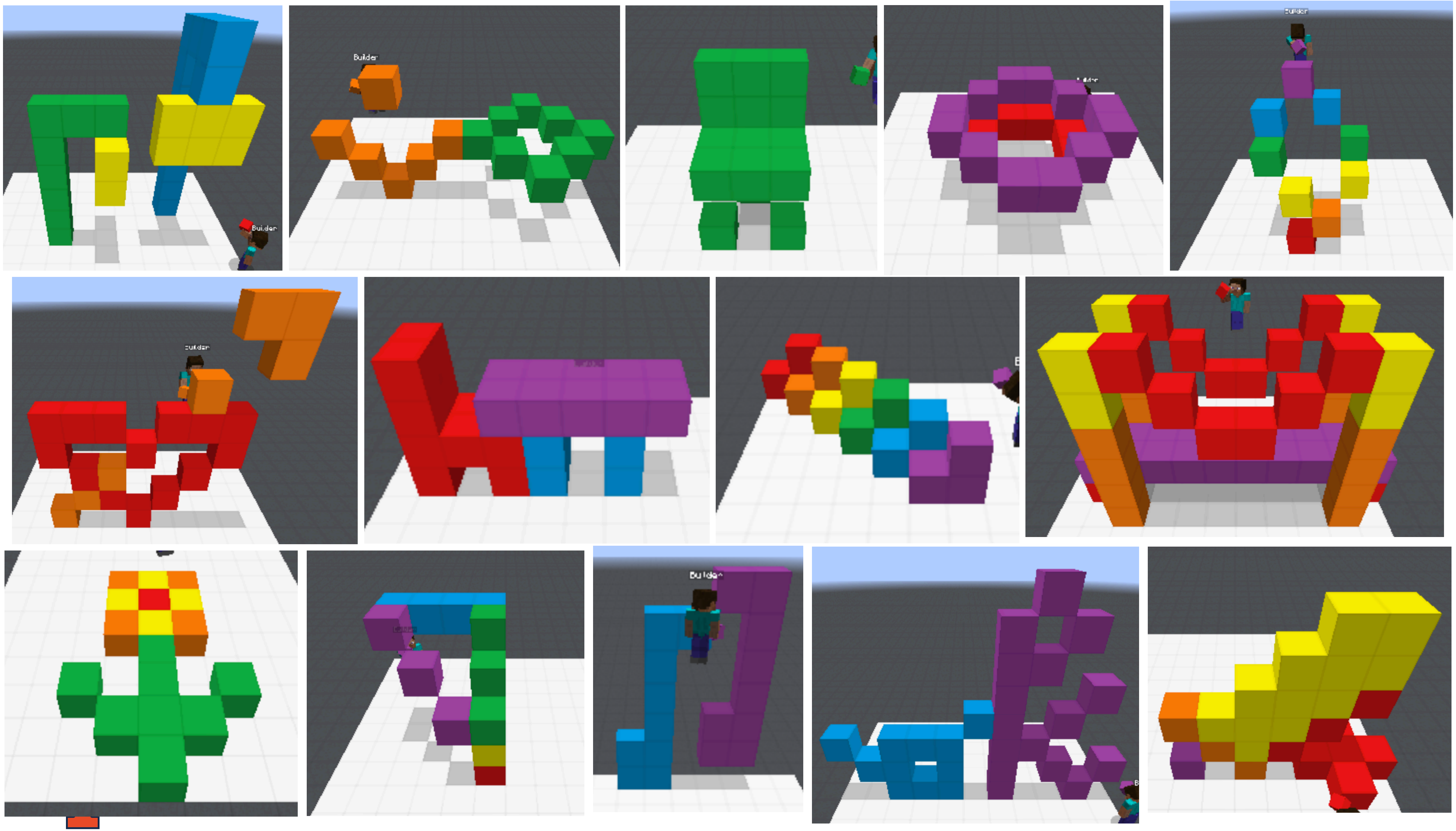


Floating blocks require temporary supports that need to be removed again


<Architect> go the middle and place an orange block two spaces to the left
<Architect> now make a staircase with 2 stairs left and 2 right with orange
<Architect> so it will look like a v
Builder
Floating blocks require temporary supports that need to be removed again


<Architect> go the middle and place an orange block two spaces to the left
<Architect> now make a staircase with 2 stairs left and 2 right with orange
<Architect> so it will look like a v
Builder
Floating blocks require temporary supports that need to be removed again

Our Dataset:
The Minecraft Dialogue Corpus



Minecraft Dialogue Corpus

(Narayan-Chen, Jayannavar & Hockenmaier, 2019)

– **150 Target structures**, split across train/test/dev
– **509 Human-human dialogues & game logs** for the Collaborative Building Task
– **15.9k Utterances** (**11.5k Architect**, 4.4k Builder)
– **6.6k Builder** action sequences

– Built on top of **Microsoft's Project Malmo**
– You can **download** our data and data collection code
– Caveat: data collection requires users to have our version of Minecraft/Malmo on their machines

How Can We Build Systems that Can Perform This Task?

How can we build systems that can perform this task?

Option 1:

Develop rich linguistic representations for this domain

Annotate the Minecraft Dialogue Corpus

Train generation and parsing models on these annotations

Develop agents that use these models

Option 2:

Train end-to-end neural models on this data

AN INTERMEDIATE LINGUISTIC REPRESENTATION: AMRS FOR DIALOGUE AND SPATIAL RELATIONS

BONN, PALMER, CAI, WRIGHT-BETTNER, LREC 2020

Bonn, Palmer, Cai, Wright-Bettner, LREC 2020

Spatial PropBank Rolesets

Rolesets do the heavy lifting in AMRs:

Multi-alias Rolesets: apply to a fixed set of synonymous spatial expressions
Roles: semantic/pragmatic roles, annotated with participants from the text
Entailments: meaning expressed by the roleset itself, no need to annotate manually

left-j
leftward-r
on_the_left-m

left-20
ARG1-SE1 entity on the left
ARG2-SE2 of what?
ARG3-ANC anchor for FoR
ARG4-AXS axis

discrete(SE1, SE2)
framework(ANC)
horizontal(AXS, ANC)

Expanded Roleset Inventory:
186 new/updated rolesets
verbs, nouns, adjectives, prepositions, adverbs, MWEs
20 new semantic/pragmatic role types

AMR Annotation

Single-sentence annotation: surface representation & frame of reference
Multi-sentence annotation: intersentential coreference & implicit arguments
Dummy AMR: maps specific spatial frameworks from dialogue onto absolute coordinate system

Now one to the left.

```
(b / be-destined-for-91
      :ARG1 (t / thing :quant 1)
      :ARG2 (s2 / space
            :ARG1-of (l / left-20
                  :ARG2 [implicit: previous block]
                  :ARG3 (c /cartesian-framework-91
                        :ARG1 (b2 / builder))))
      :time (n / now))
```

Inferred concept

Dialogue-level coreference

Frame of Reference co-refers to dummy AMR

Translation from surface 'left' → absolute form:
yaw= 88.2 → B is facing EAST → "left" = absolute NORTH

Annotation Statistics:

243 full dialogues
7255 dialogue sentences (+ 11,000 auto-generated non-dialogue construction AMRs)

How can we build agents that can perform this task?

Option 1:

Develop rich linguistic representations for this domain

Annotate the Minecraft Dialogue Corpus

Train generation and parsing models on these annotations

Develop agents that use these models

Option 2:

Train end-to-end neural models on this data

How can we build agents that can perform this task?

Option 1:

Develop rich linguistic representations for this domain

Annotate the Minecraft Dialogue Corpus

Train generation and parsing models on these annotations

Develop agents that use these models

Option 2:

Train end-to-end neural models on this data

Starting Point for Architect: Utterance Generation

Architect: Tasks and Challenges

Give clear and correct instructions in a changing environment

A. needs to identify **next steps** for B.
A. needs to **align target and build** region
A. needs to adapt to **B's current position**
A. needs to **identify mistakes** made by B.

Answer Builder's questions

Interrupt the Builder to correct mistakes

A. should **respond in real time** (no turns)

Architect Utterance Generation Task

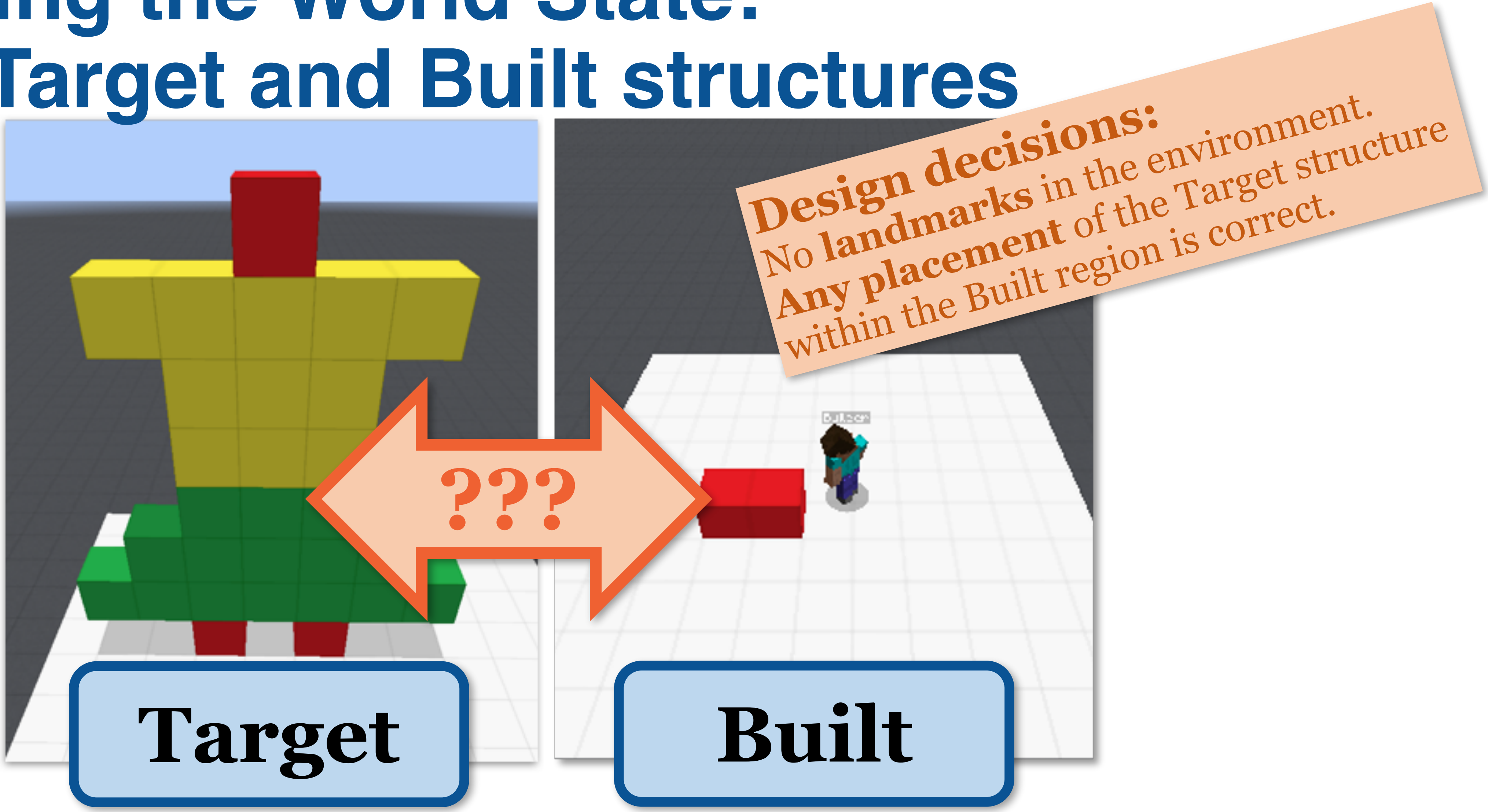
Generate a suitable Architect utterance
for a game state in a human-human game
when the human Architect said something.

Ignores **real-time** aspect
(when to speak)

Ignores **overall task** completion
(how to maintain a whole conversation)

Allows us to use supervised learning to develop **baseline models**

Modeling the World State:
Align Target and Built structures



Modeling the World State naively with Block Counters

Global Block counters (one 18-dimensional vector)

For each of the 6 colors: #blocks to be **added**, **added next**, and **removed**

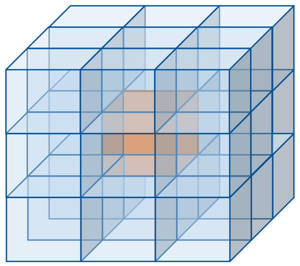
Averaged over all optimal alignments of built to target.

Add	Add Next	Remove	Add	Add Next	Remove	Add	Add Next	Remove	Add	Add Next	Remove	Add	Add Next	Remove	Add	Add Next	Remove

Local Block Counters (concatenate 27 block counters)

Separate counters for each cell in the **3×3×3 cube around the last cell** the Builder touched.

To capture the Builders' current perspective, the order of cells depends on the Builder's current position, pitch and yaw.

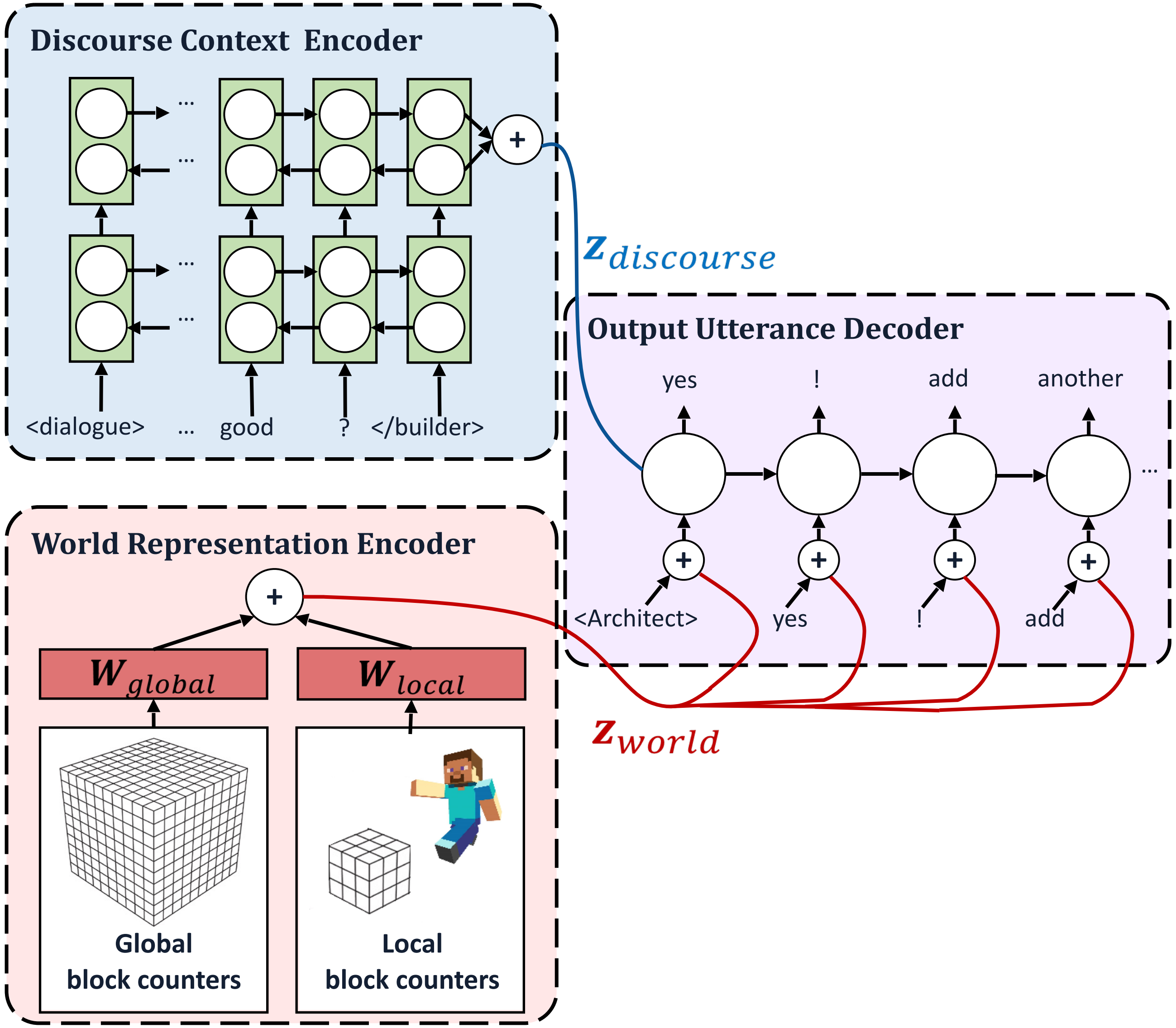


Our Model

Discourse Context encoder:
biGRU over previous dialogue with Glove embeddings

World Context Encoder:
$\mathbf{W}_{global}$: Global Block counters
$\mathbf{W}_{local}$: Local Block counters

Output utterance decoder:
Reads block counter embeddings (and last token) at each time step



Automatic Evaluation

Automatic Evaluation	BLEU-1
seq2seq	15.3
Block Counter	**15.7**

Block Counter model gives a **minor improvement in BLEU-1.**

Automatic Evaluation

Automatic Evaluation	BLEU-1	Spatial P/R
seq2seq	15.3	**9.3**/8.6
Block Counter	**15.7**	8.7/8.7

Block Counter model gives a **minor improvement in BLEU-1.**
Block Counter model has **slightly lower performance on spatial terms.**

Automatic Evaluation

Automatic Evaluation	BLEU-1	Spatial P/R	Color P/R
seq2seq	15.3	**9.3**/8.6	8.1/17.0
Block Counter	**15.7**	8.7/8.7	**14.9/28.7**

Block Counter model gives a **minor improvement in BLEU-1.**
Block Counter model has **slightly lower performance on spatial terms.**
Block Counter model has **much better precision and recall of color terms.**

Human Evaluation

How **correct** are the generated utterances
(wrt. **current game state and target**)?
Correct utterances are more likely to lead to **task completion**.

	Fully correct	Partially correct	Incorrect
Human (ceiling)	**89.0%**	**0.0%**	**0.0%**

Most human utterances are **fully correct**
(remainder: correctness can't be assessed, e.g. in chit-chat)

Human Evaluation

How **correct** are the generated utterances
(wrt. **current game state and target**)?
Correct utterances are more likely to lead to **task completion**.

	Fully correct	Partially correct	Incorrect
Human (ceiling)	**89.0%**	**0.0%**	**0.0%**
seq2seq (baseline)	14.0%	28.0%	**48.0%**

Almost half of the **baseline model**'s utterances are incorrect.

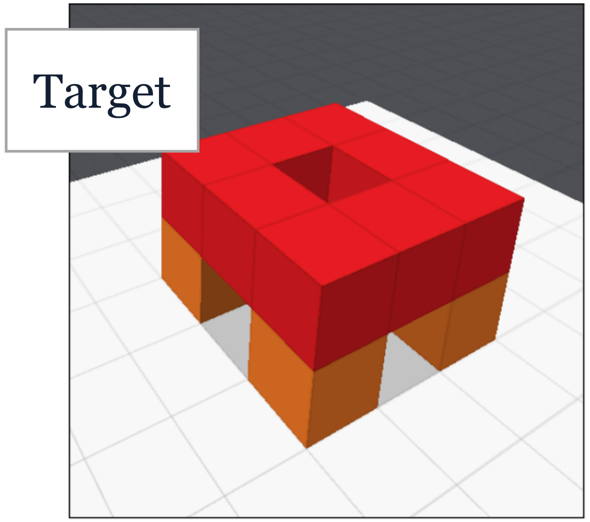
Human Evaluation

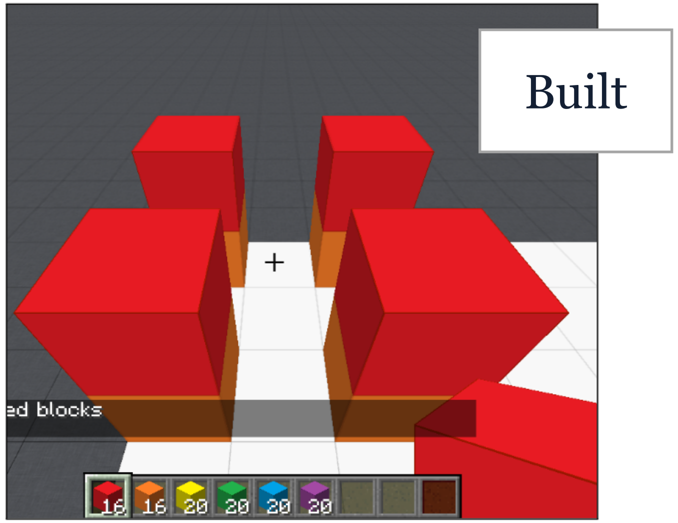
How **correct** are the generated utterances
(wrt. **current game state and target**)?
Correct utterances are more likely to lead to **task completion**.

	Fully correct	Partially correct	Incorrect
Human (ceiling)	**89.0%**	**0.0%**	**0.0%**
seq2seq (baseline)	14.0%	28.0%	48.0%
Block Counters	**25.0%**	**36.0%**	**32.0%**

The **Block Counter** Model produces **significantly more fully/partially correct utterances** and **significantly fewer incorrect ones** than the baseline (even if it is still pretty far from human performance)

What can the neural Architect do?





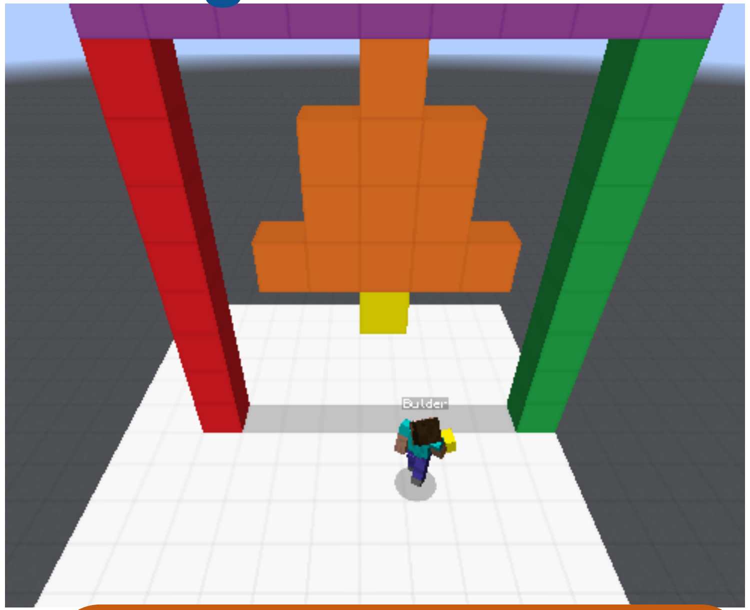
Builder has just placed the red block in the top right corner

A: *"perfect! now place a red block to the left of that"*

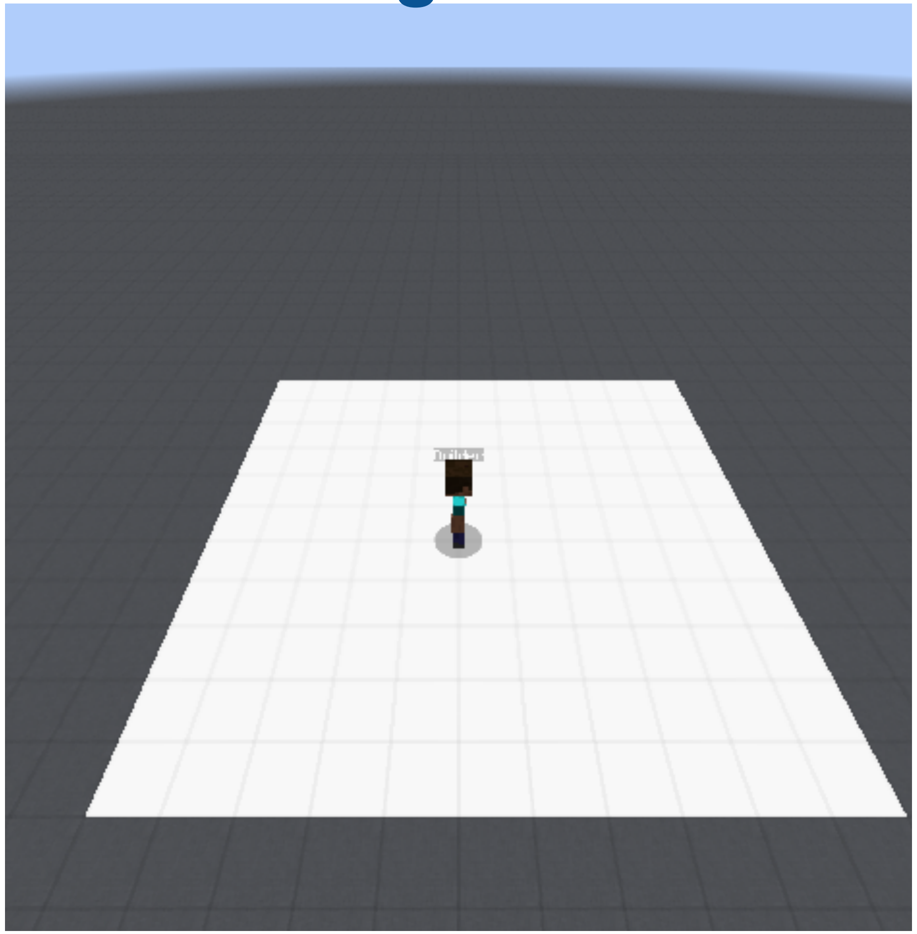
The neural architect gives natural, **fluent block-by-block instructions** that contain **color terms** and **spatial relations…**

… but it cant do much more than that

Target structure



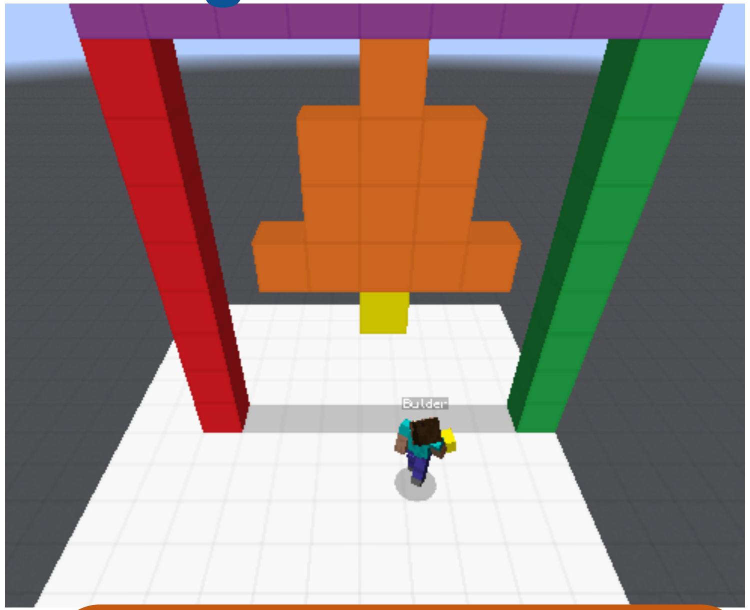
Current game state



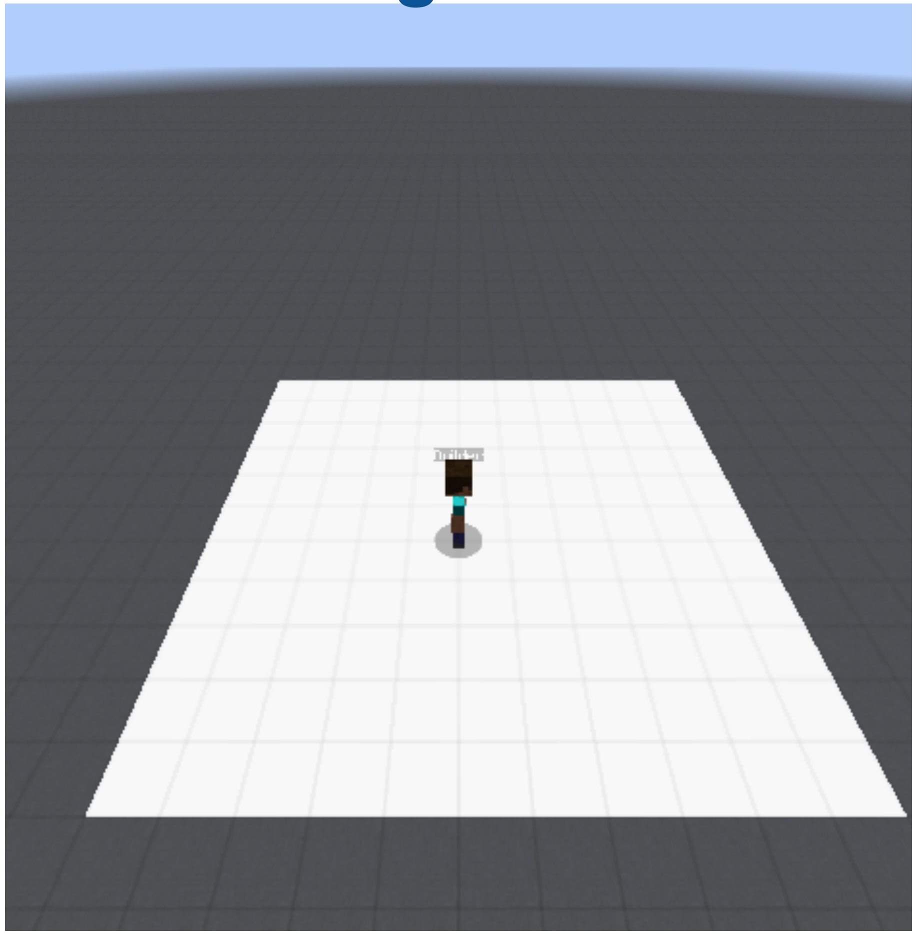
Blue: Model Architect

Model A: okay , we 'll start with a row of three red blocks , place a red block in front of you

Target structure



Current game state

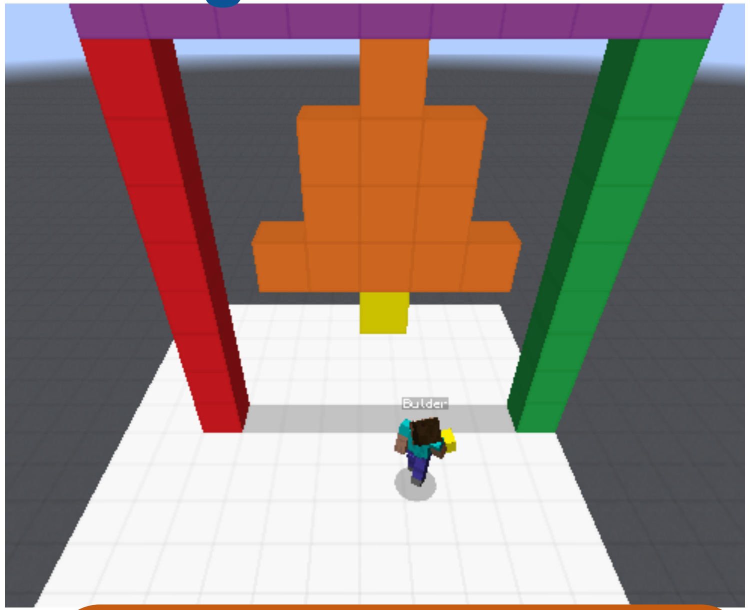


Blue: Model Architect
Red: Mistakes

Model A: okay , we 'll start with a row of three red blocks , place a red block in front of you

The block counters model has no access to **complex shapes** (*rows* vs. *towers/columns*).

Target structure



Current game state

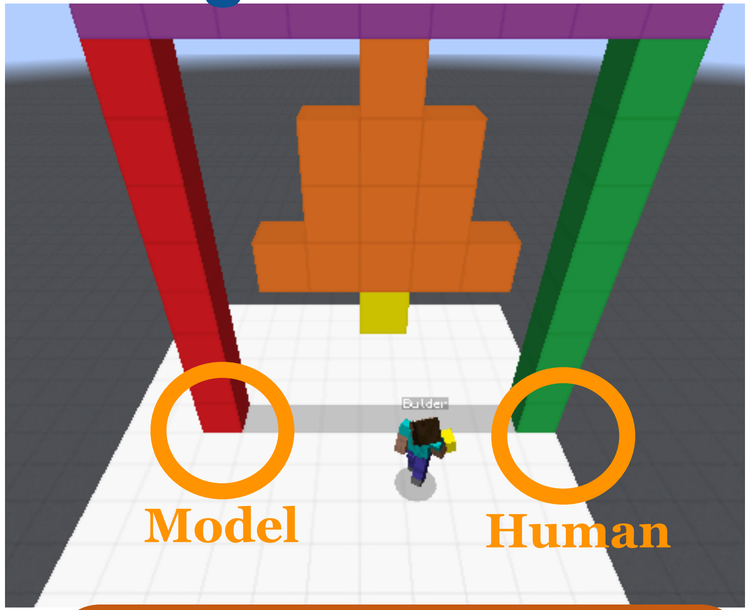
Humans give **much more complex instructions**. (“*belltower*”)

Blue: Model Architect
Red: Mistakes
Green: Human Architect

Model A: okay , we 'll start with a row of three red blocks , place a red block in front of you

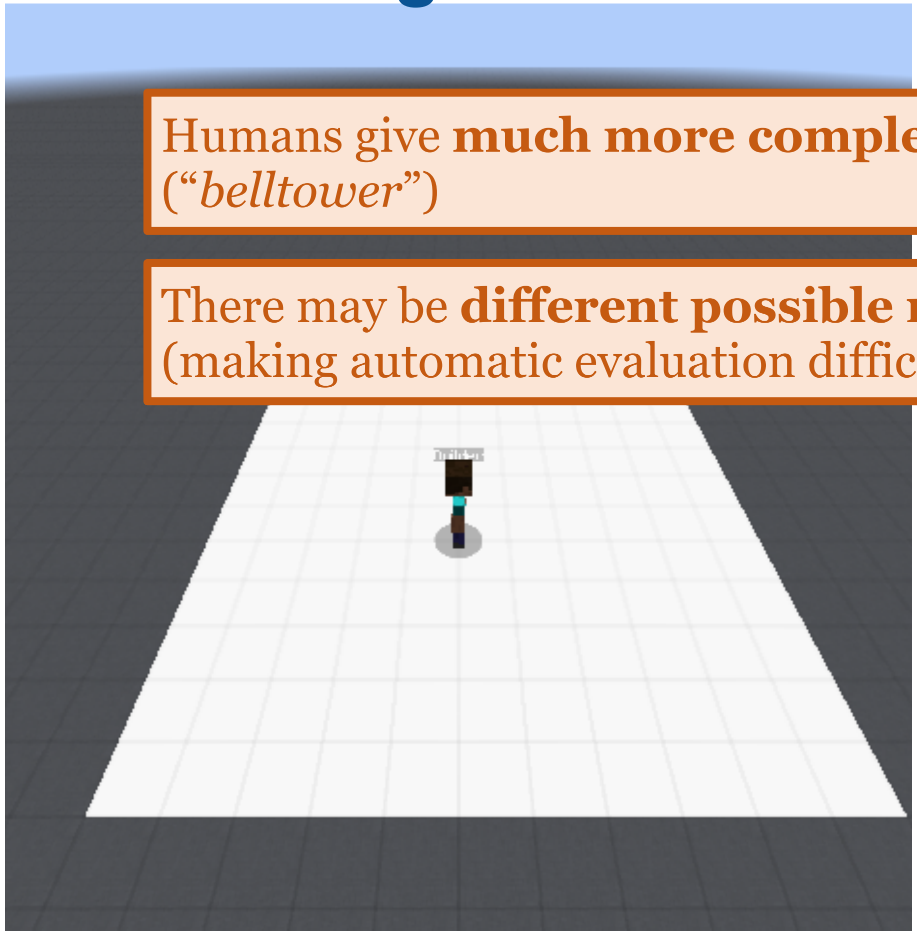
Human A: hello builder , i will tell you this. it appears we are creating a belltower . but first i will start with step by step instructions. we will start with green blocks

Target structure



Blue: Model Architect
Red: Mistakes
Green: Human Architect

Current game state



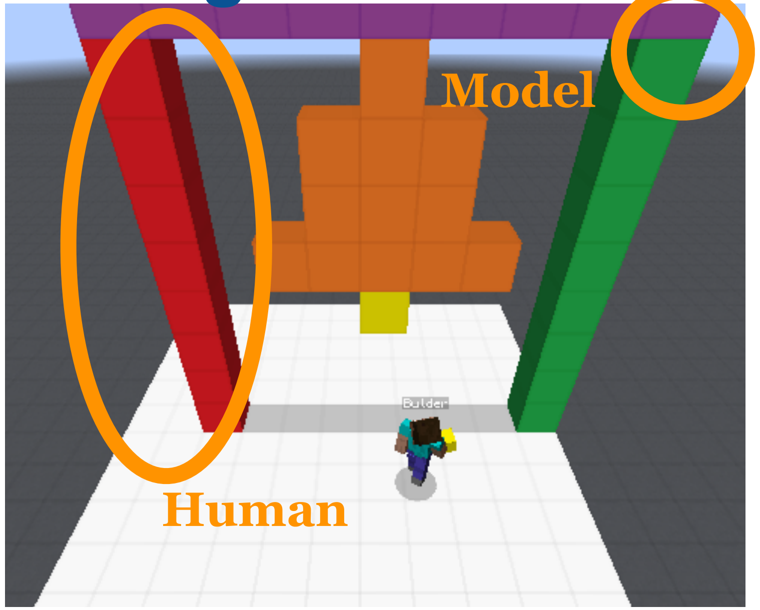
Humans give **much more complex instructions**. (“*belltower*”)

There may be **different possible next actions** (making automatic evaluation difficult)

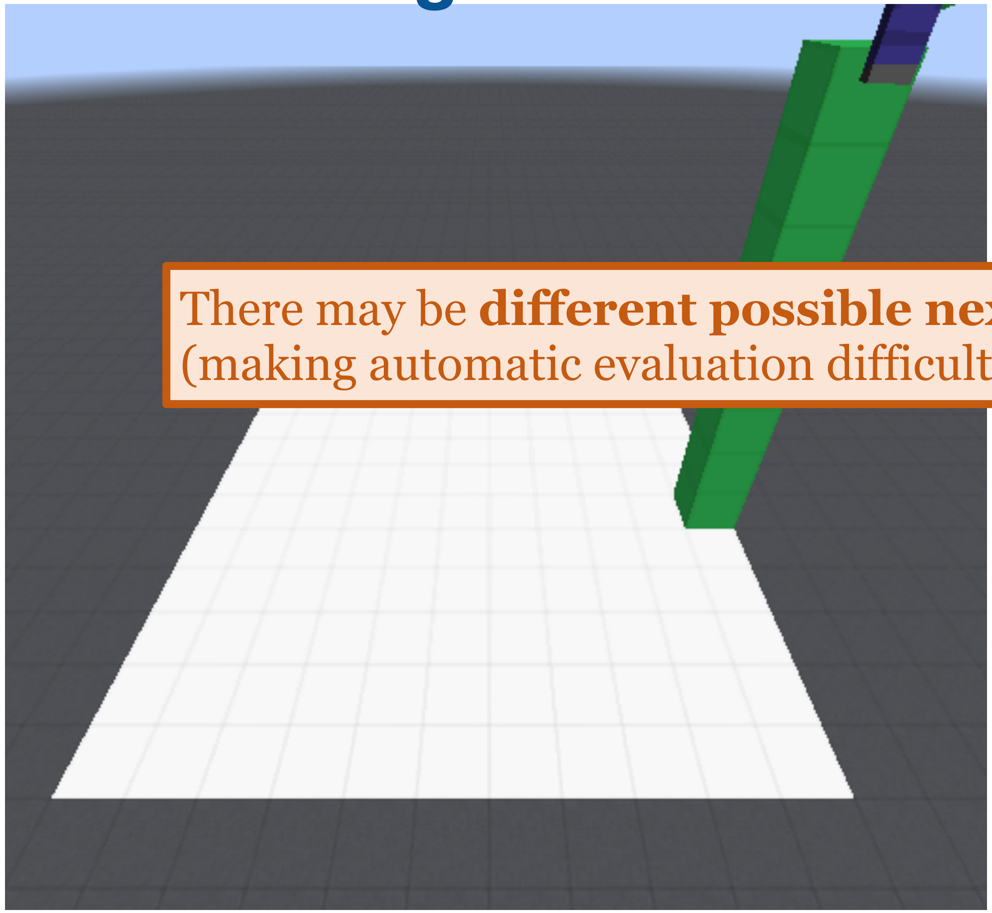
Model A: okay , we 'll start with a row of three red blocks , place a red block in front of you

Human A: hello builder , i will tell you this. it appears we are creating a belltower . but first i will start with step by step instructions. we will start with **green blocks**

Target structure



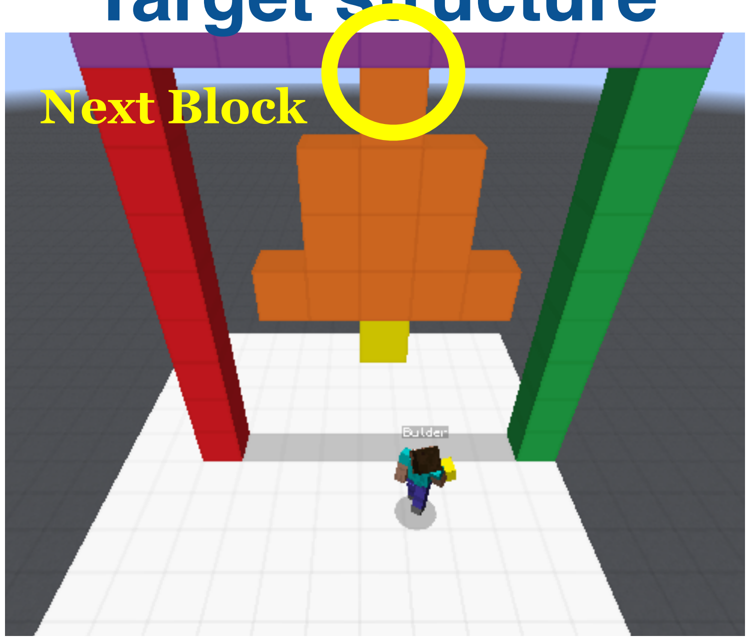
Current game state



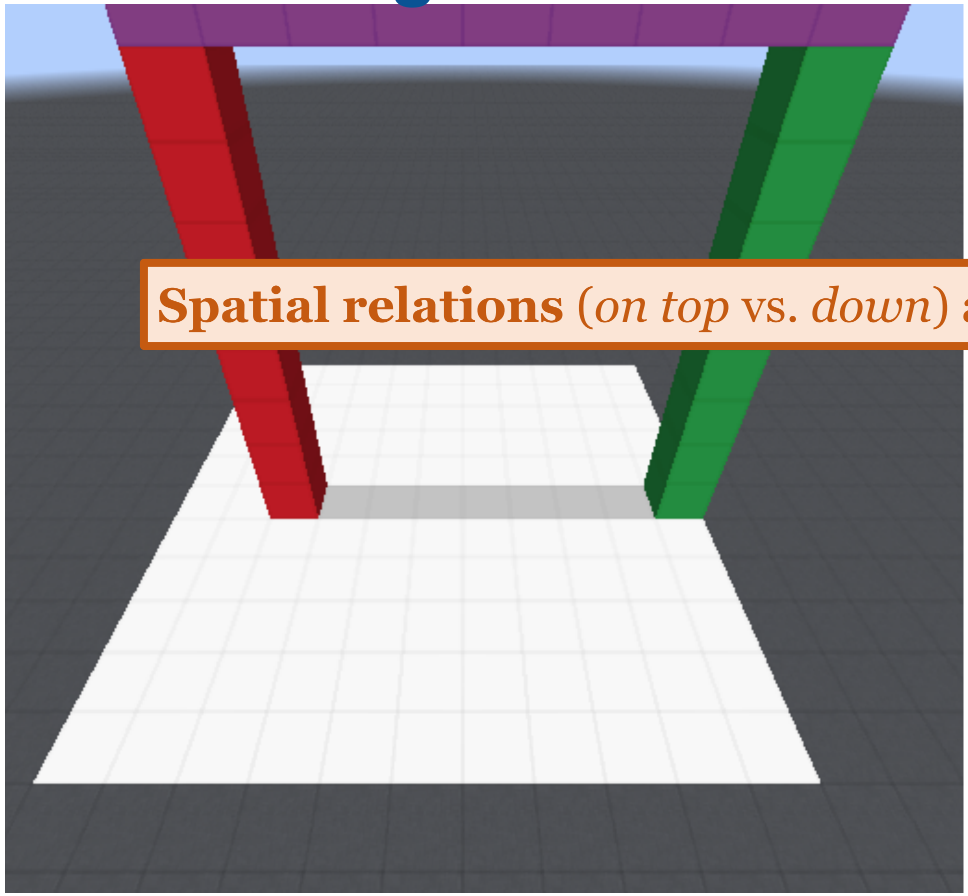
There may be **different possible next actions** (making automatic evaluation difficult)

Human B: is this good?
Human A: yes , one moment
Model A: perfect ! now put a purple block on top of that
Human A: on the 8th block to the left of the green block , please create **the same structure with red blocks**

Target structure



Current game state



Spatial relations (*on top* vs. *down*) are difficult

Human A: excellent builder

Model A: now place an orange block on top of that

Human A: now we must create the bell . please start by extending 4 orange blocks **down** from the middle purple block , as if it were hanging

Starting Point for Builder:
Action Prediction

Builder: Task and Challenges

Understand and execute instructions

B. needs to understand **descriptions of structures**
B. needs to understand **spatial relations**
B. needs to understand utterances in the **current context**

Execution: ***place and remove blocks*** *in the 11×9×11 build region*

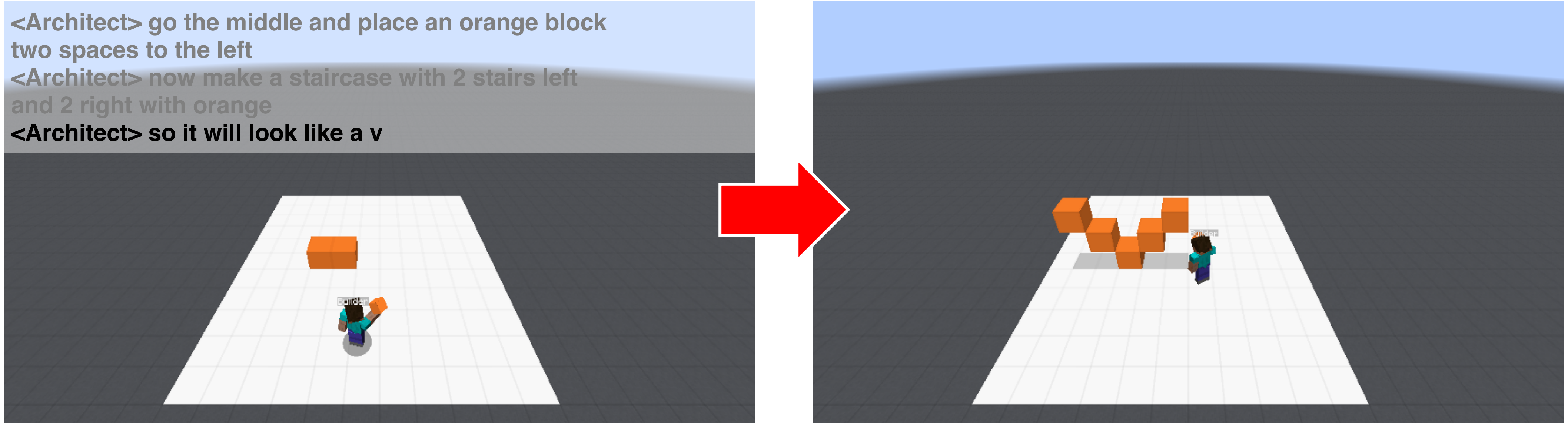
Ask clarification questions as needed

B. needs to know what information is **missing or unclear**
B. needs to know when instructions **can't be executed**

Future work:
Requires execution model

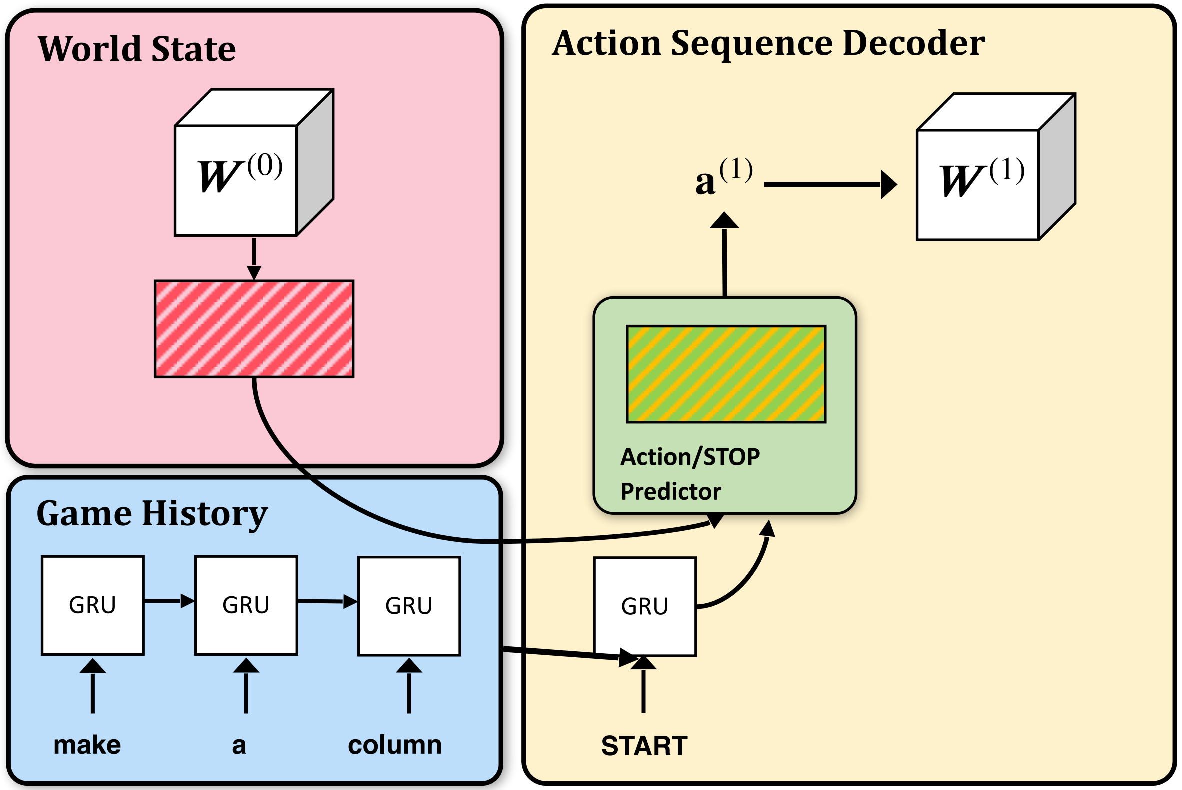
The Builder Action Prediction (BAP) Task

Predict the **sequence of actions** (block placements and/or removals) that a Builder performed at a particular point in a human-human game



Our Model

Encoder-decoder network
with GRU backbone



Inputs:

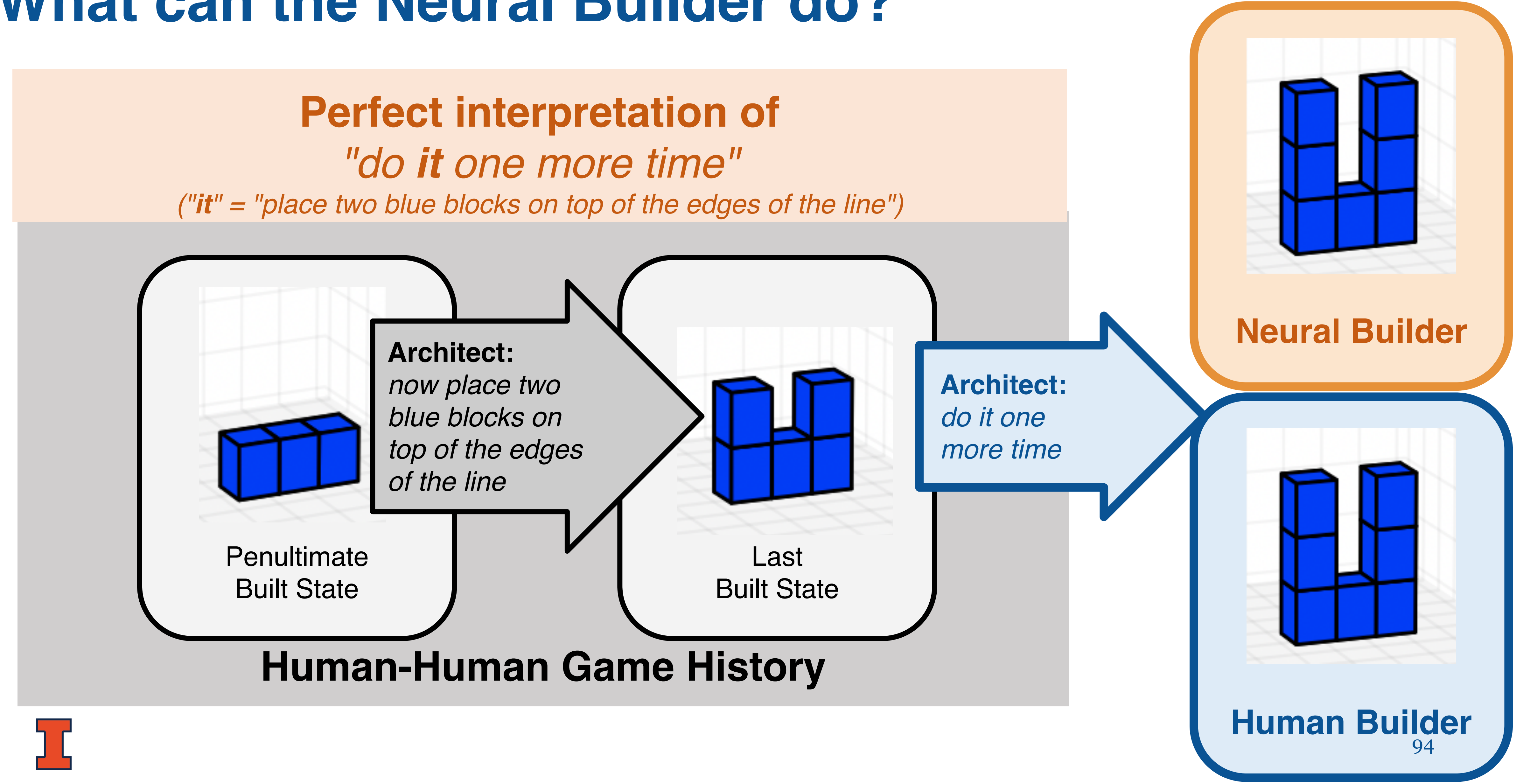
Game history up to $t = 0$

World state grid $W^{(0)}$

Predicts:

Sequence of **B** actions $\mathbf{a}^{(0)} \dots \mathbf{a}^{(t+1)}$ with $\mathbf{a}^{(0)}$ = START

What can the Neural Builder do?



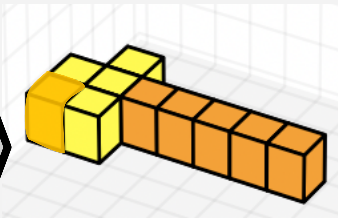
What can the Neural Builder do?

Plausible interpretation of

*"and do **the same** on **the other side**"*

*"**the same**"="the next two blocks"* *"**the other side**"=???*

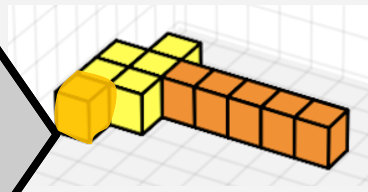
Architect:
the next two blocks will be off the corners of each of those, in the direction of the last yellow block.



Penultimate Built State

Builder:
like that, or somewhere else?

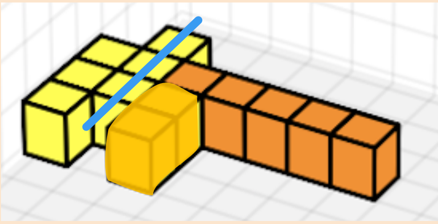
Architect:
add one more block to the end of that on your side



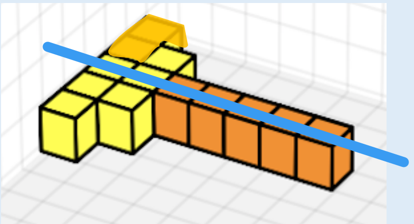
Last Built State

Human-Human Game History

Architect:
and do the same on the other side



Neural Builder



Human Builder

Plugging the Builder into Minecraft

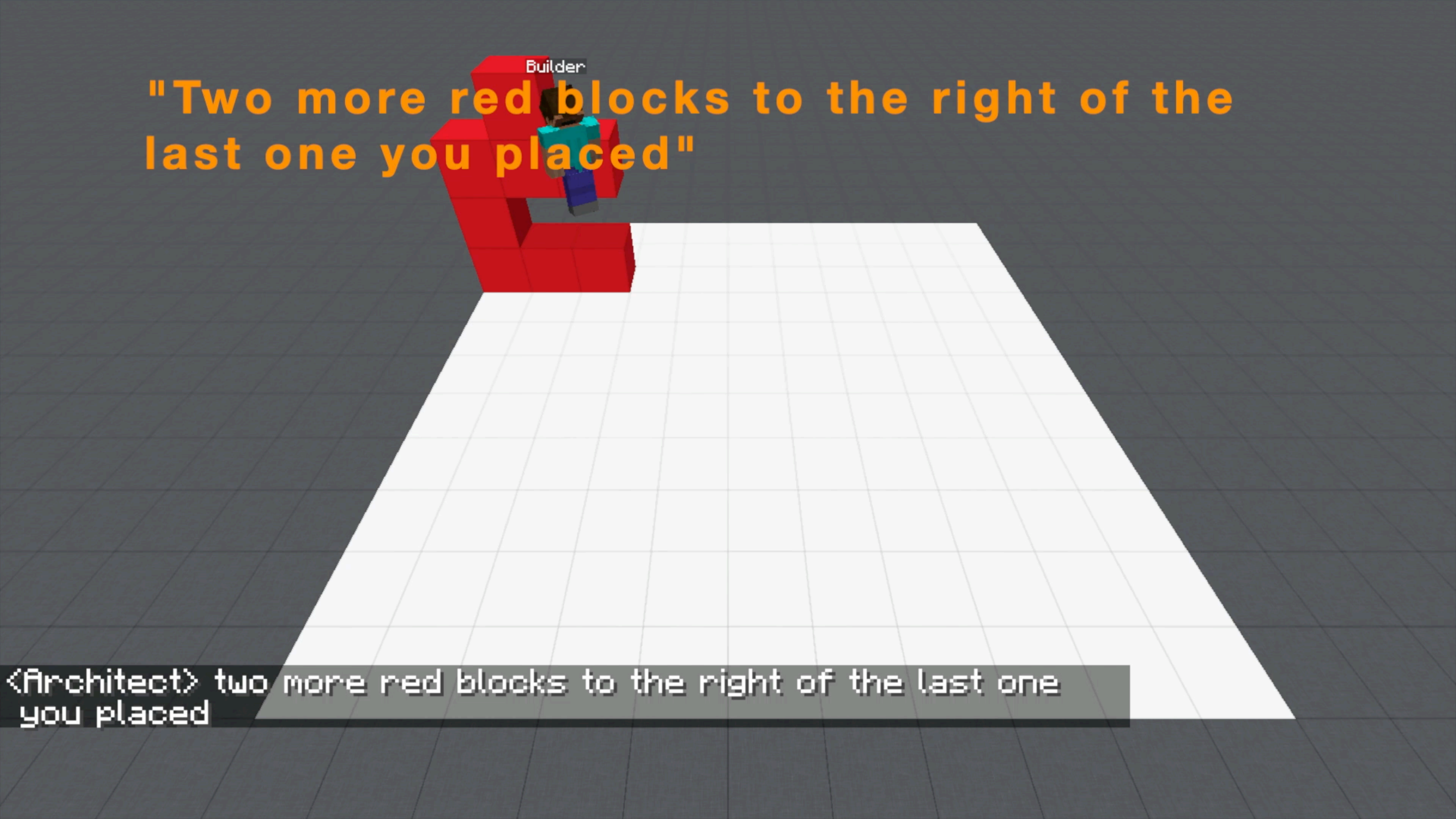
(FLOATING BLOCKS)

Plugging the Builder into Minecraft

(Spatial relations; "The Gap")

Interactive Demo:

A Human Architect gives instructions to the Neural Builder


Builder
"Two more red blocks to the right of the last one you placed"
<Architect> two more red blocks to the right of the last one you placed

What Remains To Be Done?

Fully interactive agents require further capabilities:
– Both systems need to be trained for **task completion**
– **The Builder needs to speak**, but this requires knowing **what to ask**
– Both agents need to know ***when*** **to speak**

What Remains To Be Done?

We haven't yet *solved* the tasks we started working on

– We need **higher accuracy** of instructions and executions

– We want the Architect to generate **richer, more diverse utterances**

This requires **richer models**, possibly **more data,** and other **training regimes**

– What's the role of **explicit domain knowledge?**

– Naively using **3D CNNs** as world state representations for the architect doesn't seem to work, because there is not enough supervision.

So, how can We Keep making Progress in NLP?

How do we make progress in AI/NLP?

Thanks to large amounts of raw data, industry money, GPUs, and clever neural architectures, we now have very robust models

These models work very well on many established tasks.

But we cannot work on new problems without the right datasets. New, challenging datasets can be expensive and difficult to create, especially at the scale that we need for our models to be robust.

And, we should remain mindful of the ELIZA effect!

Where will future progress come from?

Be creative — think about new tasks or domains!

Quality matters more than scale if a dataset serves as benchmark

Beware: Creating new datasets and tasks is slow

THANK YOU!