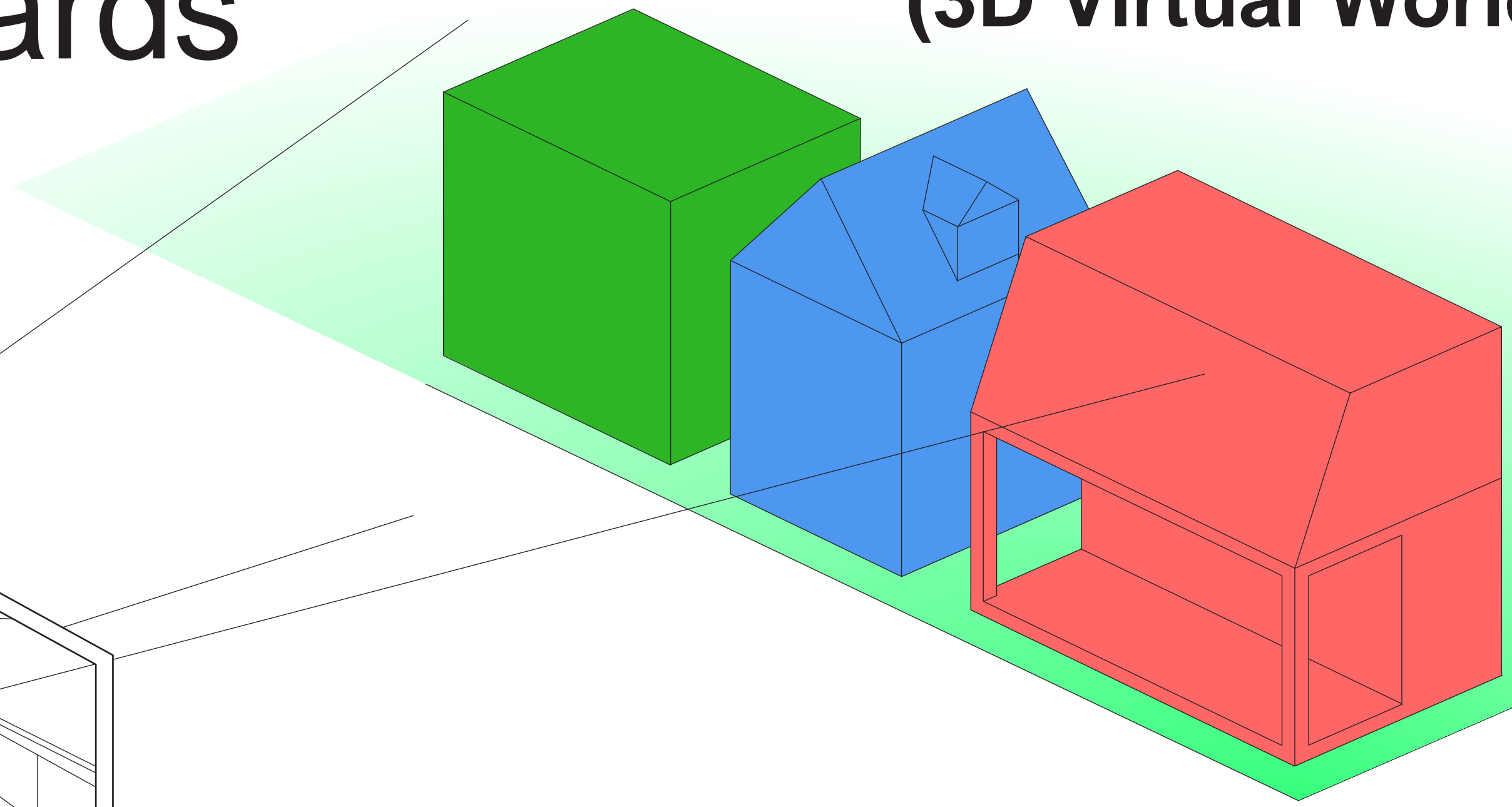


Multi-Modal Question-Answering: Questions without Keyboards

Gary Kacmarcik — Microsoft Research

Goal: Allow players in a 3D virtual world to take **virtual photographs** of objects and then use the resulting photos to interact with NPCs (non-player characters).

Scene Graph (3D Virtual World)



Object Descriptors
Generic description for each object stored as LF:

```

be 1 ({{Verb}} (.))
├── Tobj ── house 1 ({{Noun}})
│   ├── Tsub ── this 1 ({{Pron}})
│   │   └── AppInfo ── 102
│   └── Attrib ── blue 1 ({{Adj}})

```

Interaction: Player shows photo to NPC and clicks on an item to bring up a menu of dynamically generated queries.

Selecting a query from this menu, e.g.:

```

be 1 ({{Verb}} (?))
├── Tsub ── this 1 ({{Pron}})
│   └── AppInfo ── 102
└── Lnom ── what 1 ({{Pron}})

```

“What is this?”

selects the set of nodes in the KB that correspond to that object.

Treelets containing the selected nodes are combined to create:

```

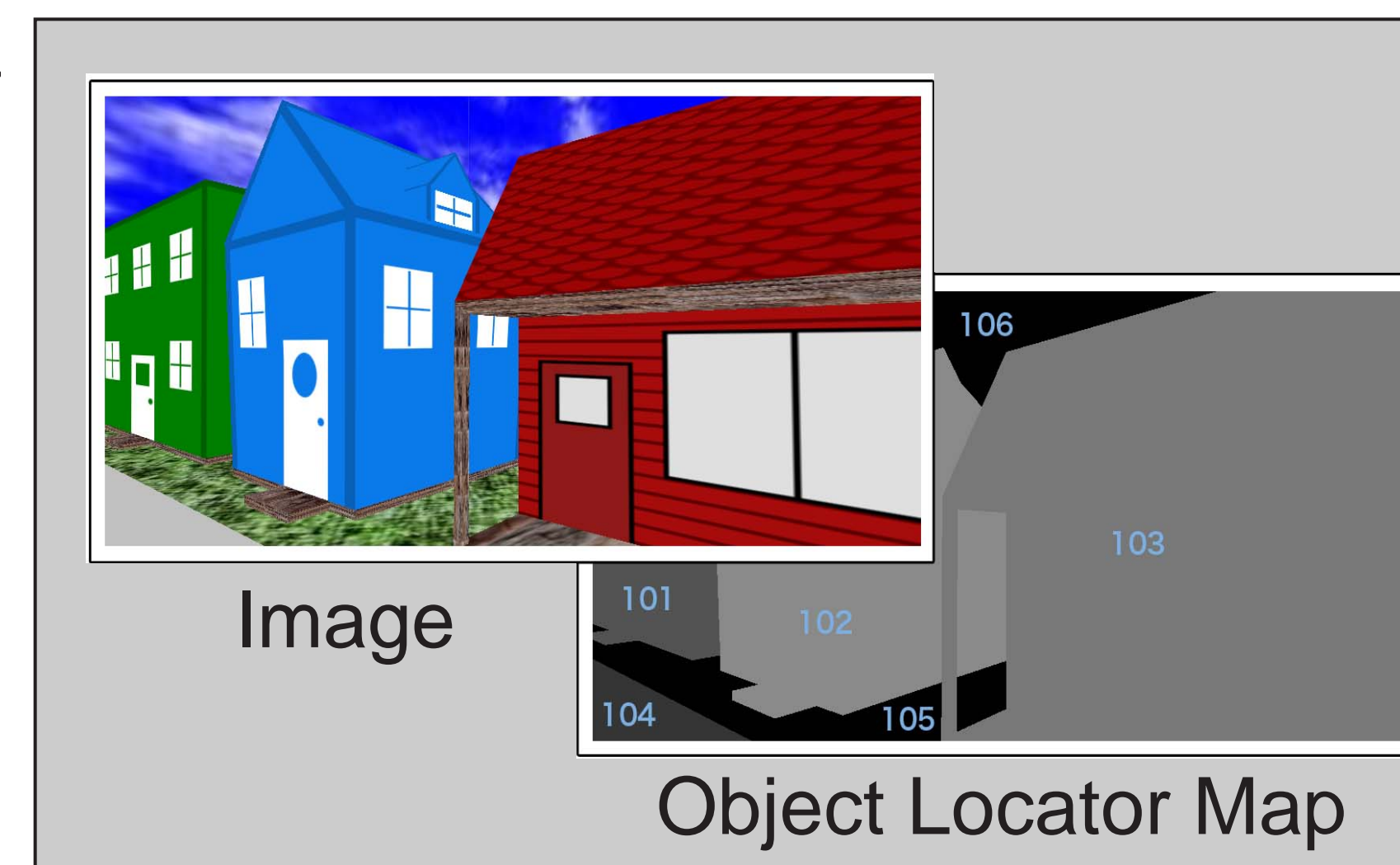
be 1 ({{Verb}} (.))
├── Tobj ── house 1 ({{Noun}})
│   ├── Tsub ── that 1 ({{Pron}} <1>)
│   │   └── AppInfo ── 102
│   └── Possr ── Mr._Jones 1 ({{Noun}})
│       └── AppInfo ── 23

```

“That is Mr. Jones’ house.”

From which we generate the final response.

Virtual Photo



+

Photo KB
Generic description of objects for use when the NPC’s KB is lacking specific knowledge

←

NPC KB
NPC knowledge and info required for deictic references.

Combined KB

```

be 1 ({{Verb}} (.))
├── Tobj ── house 1 ({{Noun}})
│   ├── Tsub ── this 1 ({{Pron}} <1>)
│   │   └── AppInfo ── 102
│   └── Possr ── John 1 ({{Noun}})
│       └── AppInfo ── 23
be 1 ({{Verb}} (.))
├── Tobj ── name 1 ({{Noun}})
│   ├── Tsub ── Mr._Jones 1 ({{Noun}} <2>)
│   │   └── Possr ── I1 ({{Pron}} <1>)
│   └── for ── John 1 ({{Noun}})
│       └── AppInfo ── 23

```

(and many more...)