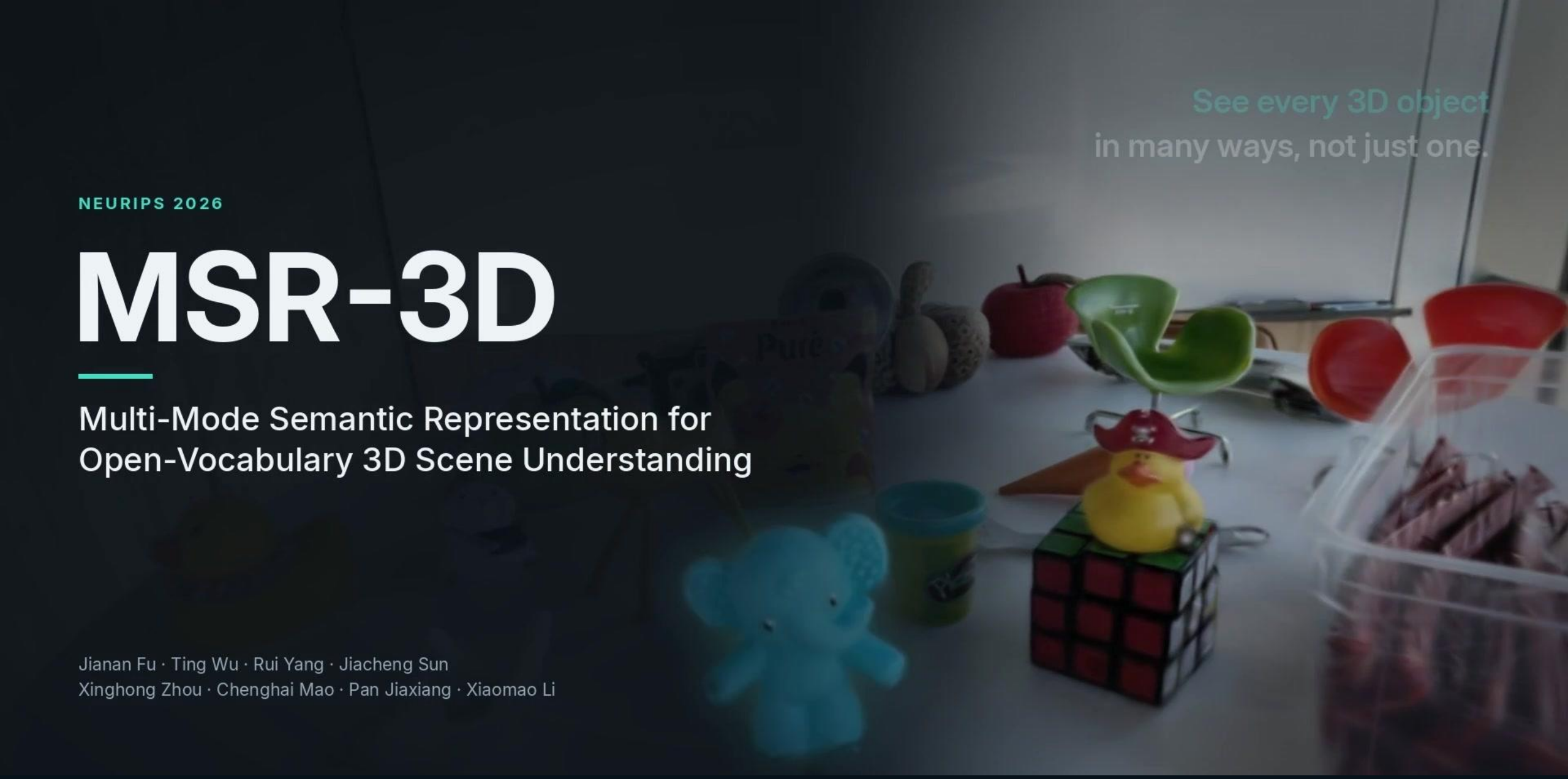
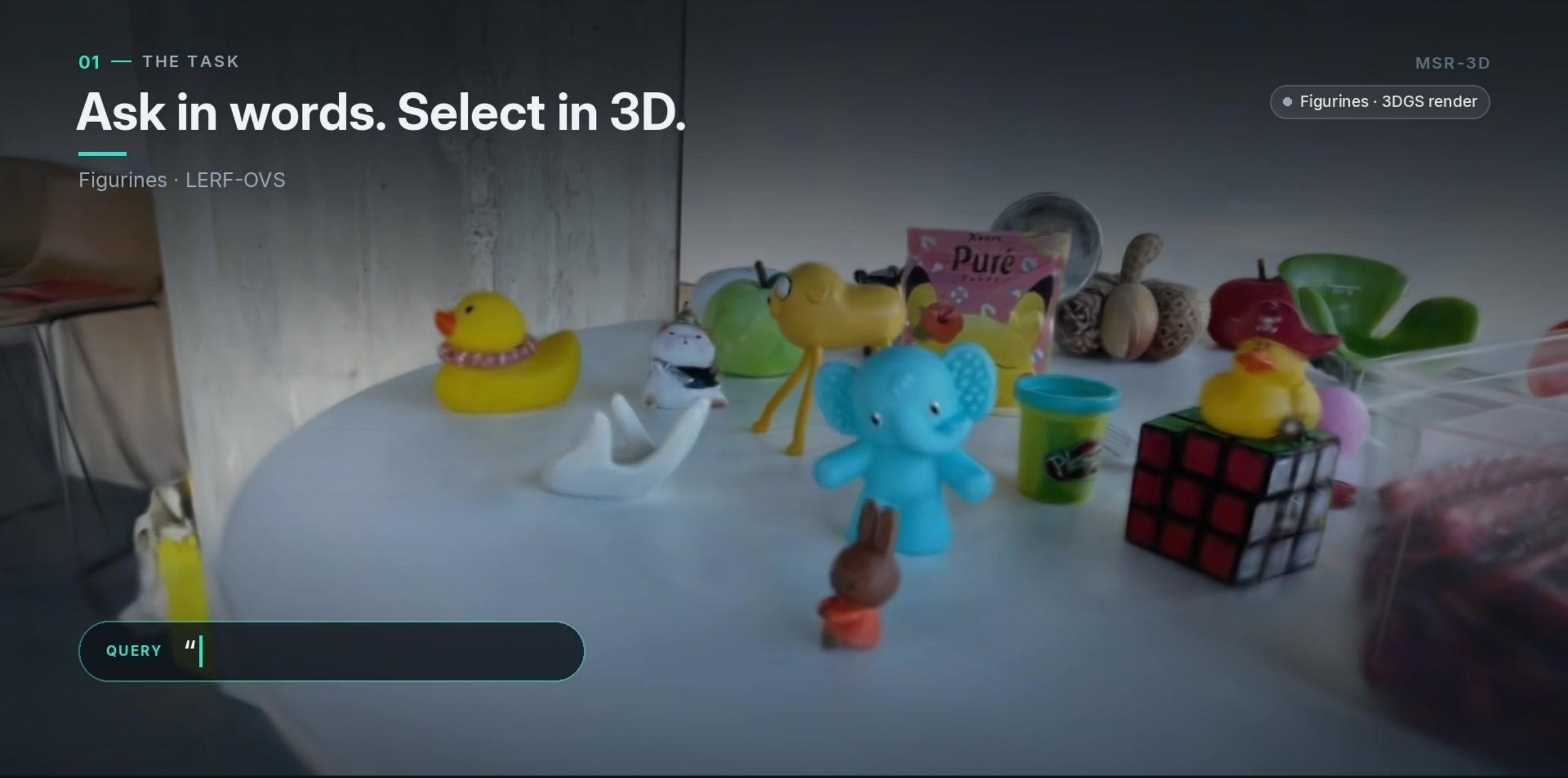


NEURIPS 2026

MSR-3D

Multi-Mode Semantic Representation for Open-Vocabulary 3D Scene Understanding

Jianan Fu · Ting Wu · Rui Yang · Jiacheng Sun
Xinghong Zhou · Chenghai Mao · Pan Jiaxiang · Xiaomao Li



Ask in words. Select in 3D.

One object. Different evidence.

Figurines · 3DGS render + training images

The same instance seen from three training cameras

VIEW A | VIEW B · CLOSE | VIEW C · OCCLUDED

TRAINING IMAGE · #00058



- 2D SAM mask (this image)
- 3D instance_0025 projected into the render

One object. Different evidence.

What does one descriptor lose?

Common practice: fuse all views into one descriptor



15 of the many training views of the toy elephant



What does one descriptor lose?

Keep structure in the evidence

MSR-3D: several modes + statistical attributes



15 of the many training views of the toy elephant



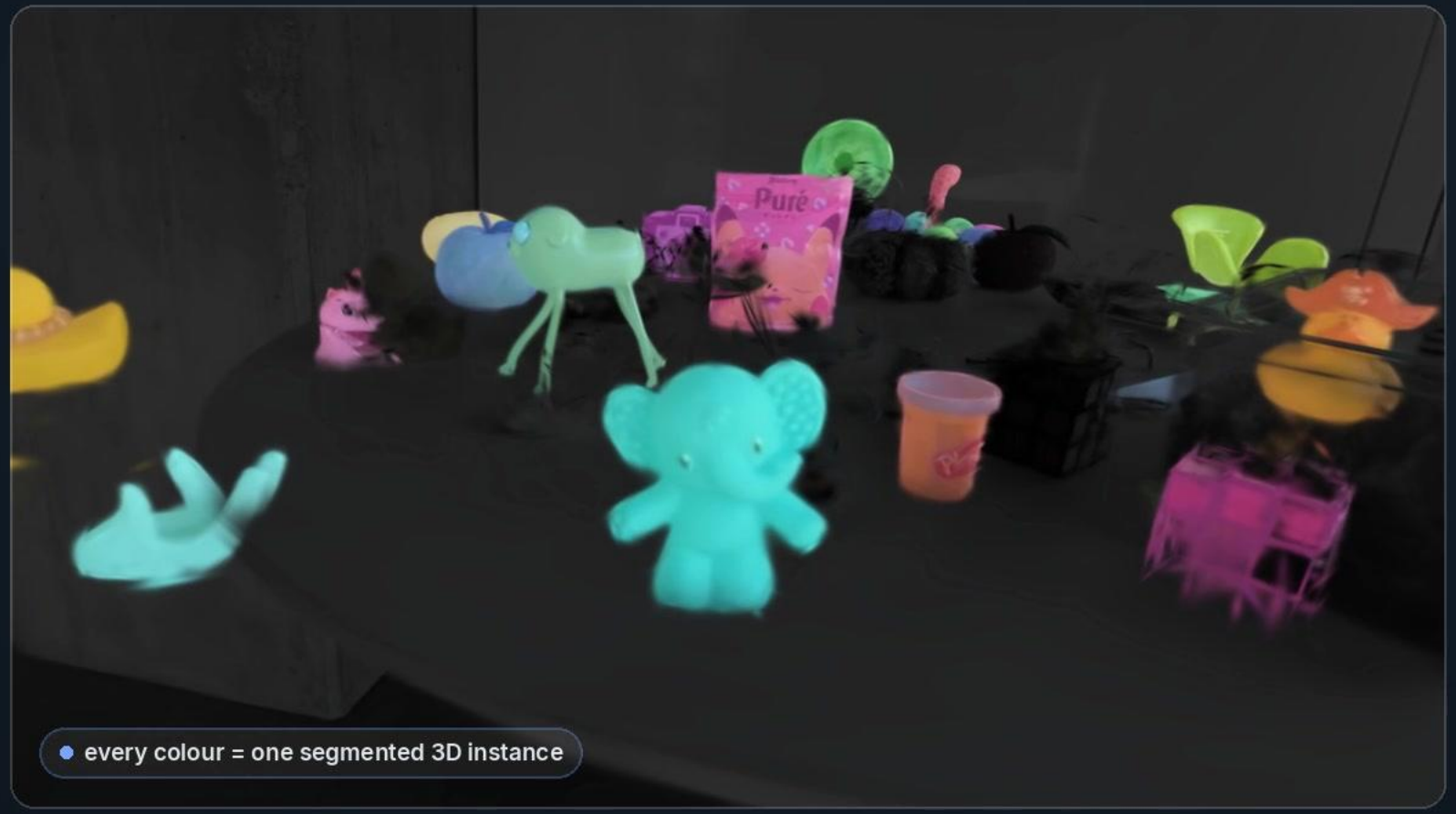


Keep structure in the evidence

Start from segmented 3D instances

Figurines · 3DGS instance labels

3DGS scene with per-Gaussian instance labels



Start from segmented 3D instances

Two mask sources, one shared encoder

2D-segmentation (SAM) and 3D-segmentation (rendered instance) observations

3D INSTANCE · ITS OWN GAUSSIANS



free of occluders, from any viewpoint

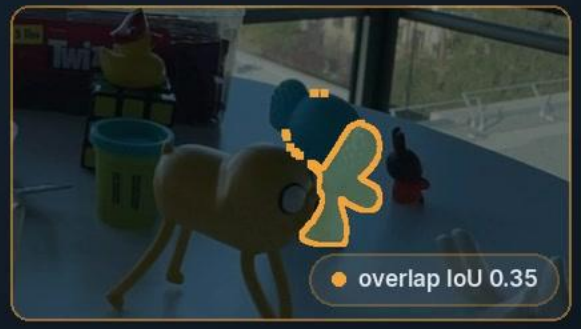
project

3D-SEGMENTATION · RENDERED



covers the whole instance

2D-SEGMENTATION · SAM



sees only a visible fragment

Two mask sources, one shared encoder

An observation pool for each instance

Four training views of instance_0025

2D SAM mask

3D rendered mask

TRAINING VIEW #00023



2D–3D IoU 0.95

TRAINING VIEW #00058



2D–3D IoU 0.97

TRAINING VIEW #00152



2D–3D IoU 0.96

TRAINING VIEW #00170



2D–3D IoU 0.35

occluded · SAM sees a fragment

OBSERVATION POOL · instance_0025

An observation pool for each instance

Preserve several semantic modes

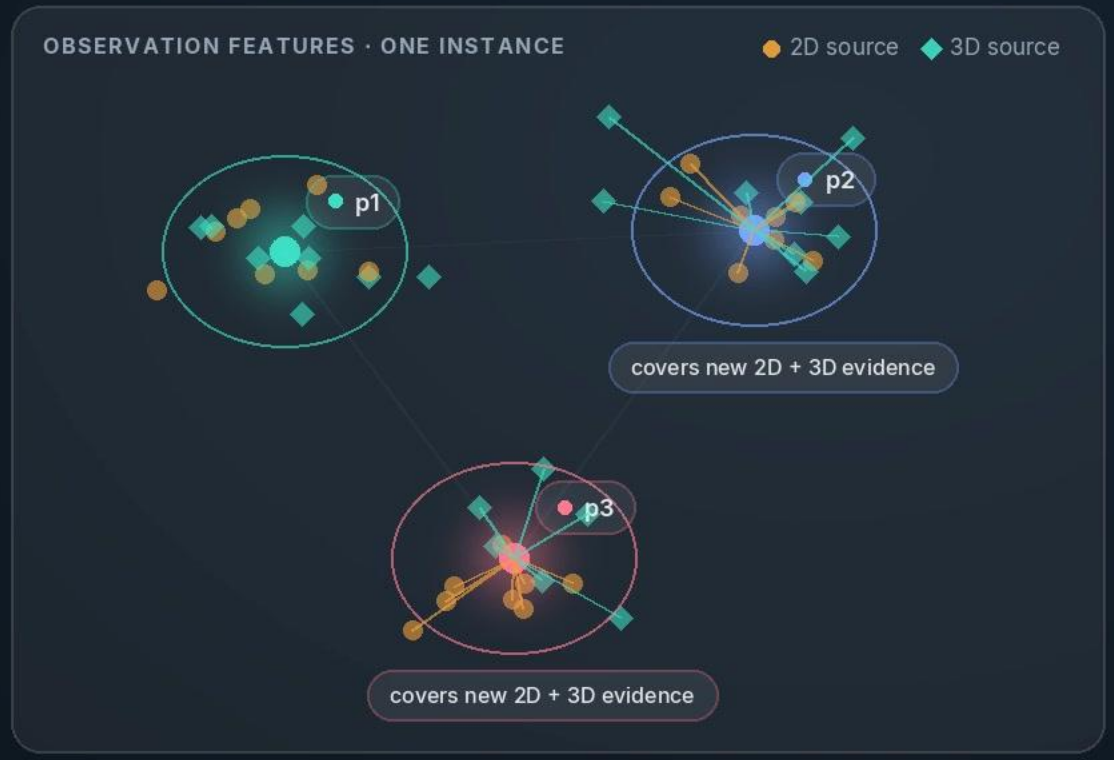
Prototypes summarize groups of nearby features



Preserve several semantic modes

Three query-independent rules

Geometric reference · source-balanced coverage · diversity



RULE 1

Geometric reference

p1 = mean of the 5 most complete observations

RULE 2

Source-balanced coverage

cover what is not yet represented · 2D and 3D weigh equally

RULE 3

Diversity

prototypes stay apart, not duplicated

Three query-independent rules

A feature alone is not the whole mode

Prototype = representative feature + statistical attributes



Attributes tell the matcher how reliable each mode is: how many observations support it, whether both sources do, and how compact and distinct it is.

A feature alone is not the whole mode

Summarize the instance, too

Prototype count · dominant support · source agreement



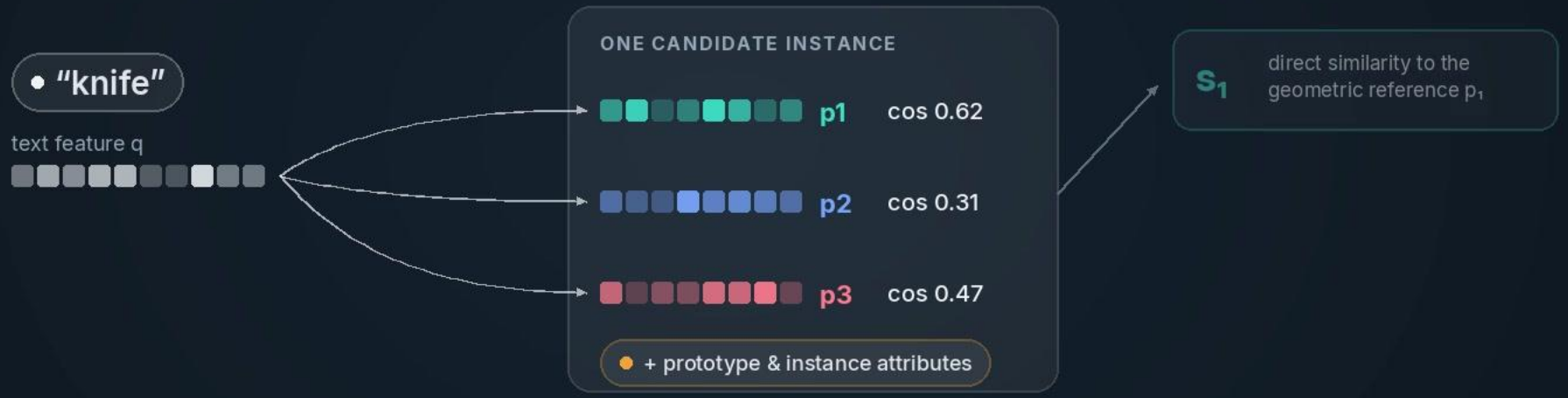
INSTANCE-LEVEL ATTRIBUTES

K_2	r^{dom}	a^{src}
prototype count	dominant support	source agreement

Summarize the instance, too

Ask the representation a question

Text–prototype similarity + statistical attributes → three-dimensional score

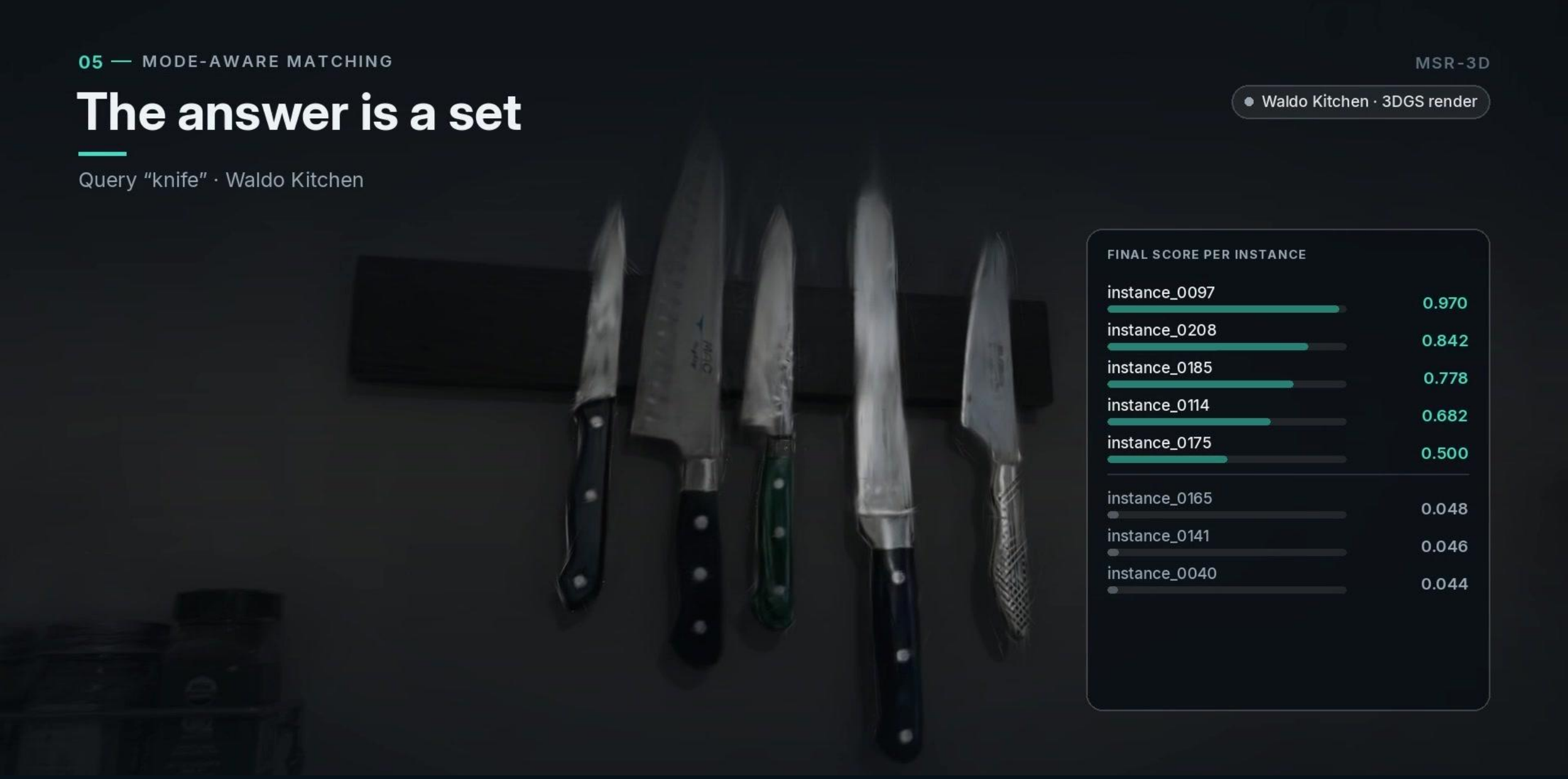


Ask the representation a question

The answer is a set

Query “knife” · Waldo Kitchen

Waldo Kitchen · 3DGS render



FINAL SCORE PER INSTANCE

instance_0097	0.970
instance_0208	0.842
instance_0185	0.778
instance_0114	0.682
instance_0175	0.500
instance_0165	0.048
instance_0141	0.046
instance_0040	0.044

The answer is a set

06 — SELECTED EXAMPLES

“nori”

Ramen · LERF-OVS

MSR-3D

Ramen · 3DGS render

TRAINING IMAGE · SAME SCENE



instance_0096 · 1 instance selected

QUERY “nori”

“nori”

06 — SELECTED EXAMPLES

“orange cat”

Bench · 3D-OVS

MSR-3D

Bench · 3DGS render

TRAINING IMAGE · SAME SCENE



instance_0024 · 1 instance selected

QUERY “orange cat”

“orange cat”

06 — SELECTED EXAMPLES

“dinosaur”

Room · 3D-OVS

MSR-3D

Room · 3DGS render

instance_0031 · 1 instance selected



QUERY “dinosaur”

“dinosaur”

Clear gains on LERF-OVS

Results from the paper

LERF67 · four scenes · scene-averaged

Scene-averaged scores (%), higher is better · grey = 8 compared methods



Clear gains on LERF-OVS

Unseen query strings

Results from the paper

LERF45 · three target scenes · queries unseen in training

Scene-averaged scores (%), higher is better · grey = 8 compared methods



Unseen query strings

Transfer to ten 3D-OVS scenes

Results from the paper

3DOVS61 · ten scenes · 61 scene–query entries

Scene-averaged scores (%), higher is better · grey = 8 compared methods



Transfer to ten 3D-OVS scenes

Retain modes. Integrate evidence.

MSR-3D · Open-Vocabulary 3D Scene Understanding

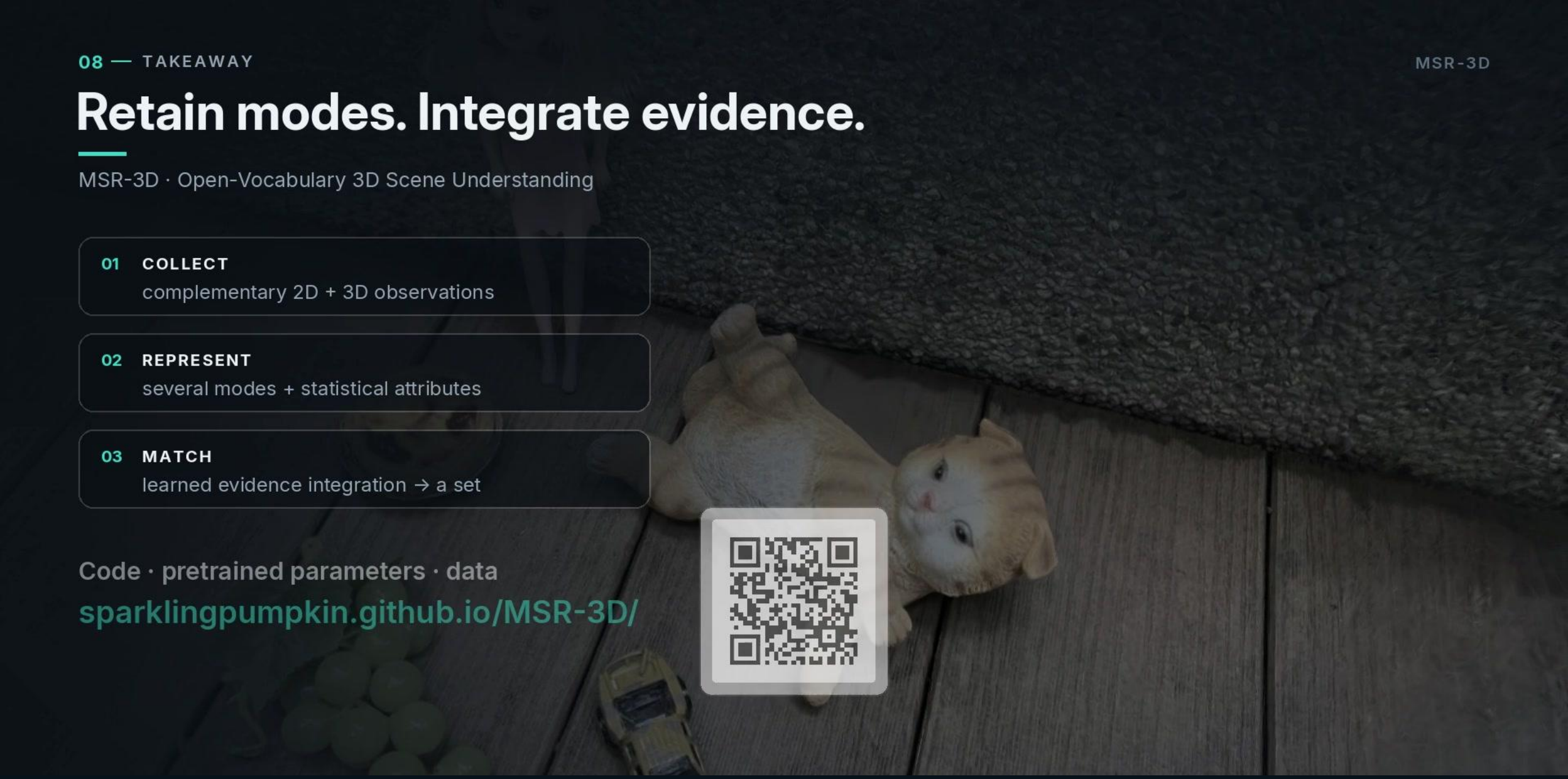
01 COLLECT
complementary 2D + 3D observations

02 REPRESENT
several modes + statistical attributes

03 MATCH
learned evidence integration → a set

Code · pretrained parameters · data

sparklingpumpkin.github.io/MSR-3D/



Retain modes. Integrate evidence.