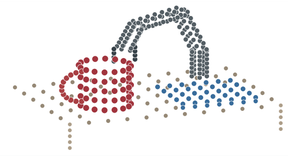
One world context, many views

Text
The blue book is on the table.

Image

Same world context

3D Point Cloud



Robot state VLA only

$$q = [q_1, q_2, \ldots]$$

Task input x

VLM Text + Image
What is beside the blue book?

VLA Text + Image + Robot state
Pick up the red mug.

Existing multimodal CoT

Encode, then concatenate

Image → image enc.
3D → 3d enc.
State → state enc.

Concat

Reasoning chain

step 1 · step 2 · step 3 · …

modality-switching gap

Uni-LaDiR (Ours)

Encode into one reasoning space

Image → image enc.
3D → 3d enc.
State → state enc.

Unified Encoder (LLM)

Reasoning chain

Diffusion

step 1 · step 2 · step 3 · …

Unified latent space

Output y

Task-dependent

VLM · Text answer

A red mug.

VLA · Action sequence

$$[a_1, a_2, \ldots]$$