Input data	Task-aware 2D Affordance Detection	3D Affordance Refinement	Output

Input data

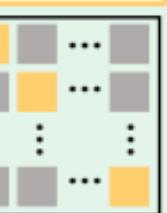task: Open the bottom closet drawer between the door and the cabinet.

video frames $\mathcal{V}$

point clouds $\mathcal{PC}$

Task-aware 2D Affordance Detection

Affordance Concept Extraction

$\mathcal{A}$ → text encoder

vision encoder

Affordance-Weighted Task Frame Selection

$\mathcal{A}$ → Double Check Mechanism

$\mathcal{V}^*$

$\mathcal{V}^*$ ↓ ↓ $\mathcal{P}^*$
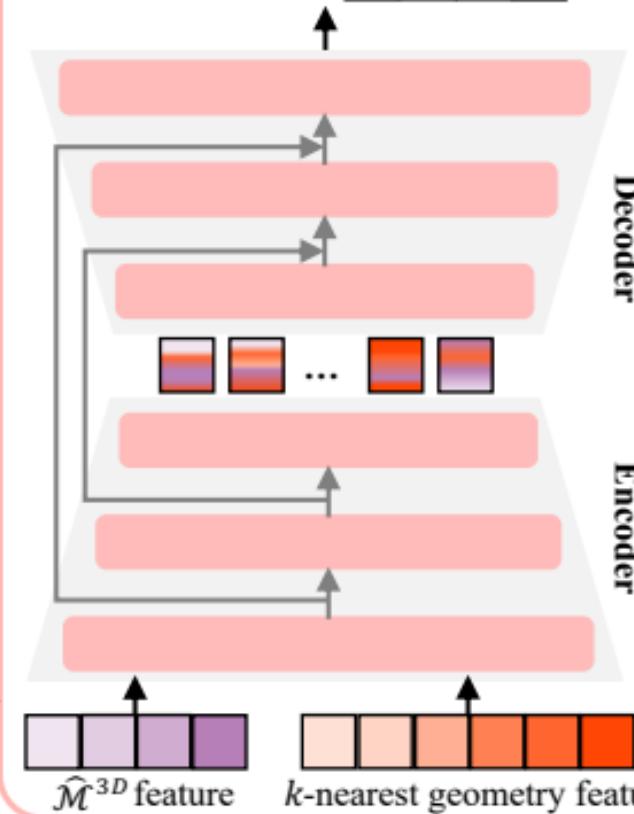
Initialization with Geometric Constraints

$\mathcal{M}^{2D}$ ⊕ $K [R|t]$ → $\hat{\mathcal{M}}^{3D}$

3D Affordance Refinement

$\mathcal{M}^{3D}$

Point Trans. Decoder

Point Trans. Encoder

$\hat{\mathcal{M}}^{3D}$ feature

k-nearest geometry feature