20M Image-Depth Pairs

Sparse Sampling

Sparse Metric Prompts

RGB Image Input

ViT Block
ViT Block
⋮
ViT Block
ViT Block

DPT Decoder
Fusion
DPT Decoder
Fusion

Metric Depth Map

(i) Recontructed 3D Data

3D Point Cloud
Connections
Frame 1
Frame f
Frame F

Camera A
Light Pattern
Object
Light Projector
Camera B

(ii) Captured 3D Data

(iii) Rendered 3D Data

(a-d) Depth Estimation, Completion，Super-Resolution, Radar–Camera Fusion

(e) Single/Multi-view 3D Reconstruction

10.8m

(f) Camera Calibration

(g) Robot Prompt Action

[x, y, z, roll, pitch, yaw, gripper]

(h) Spatial Reasoning

Human: These are frames of a video, what is the **direct distance** between the **window and the table** (in meters)?
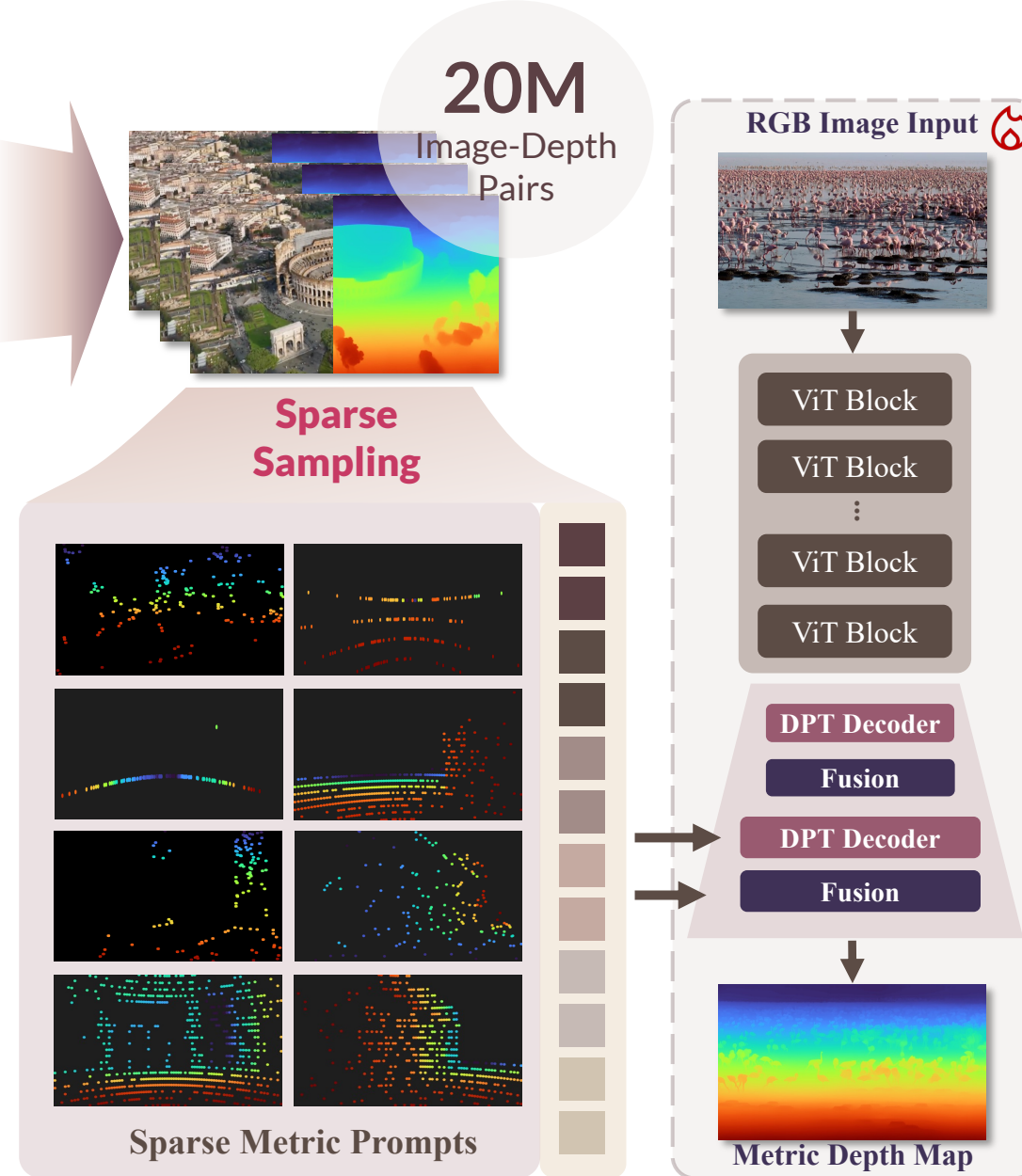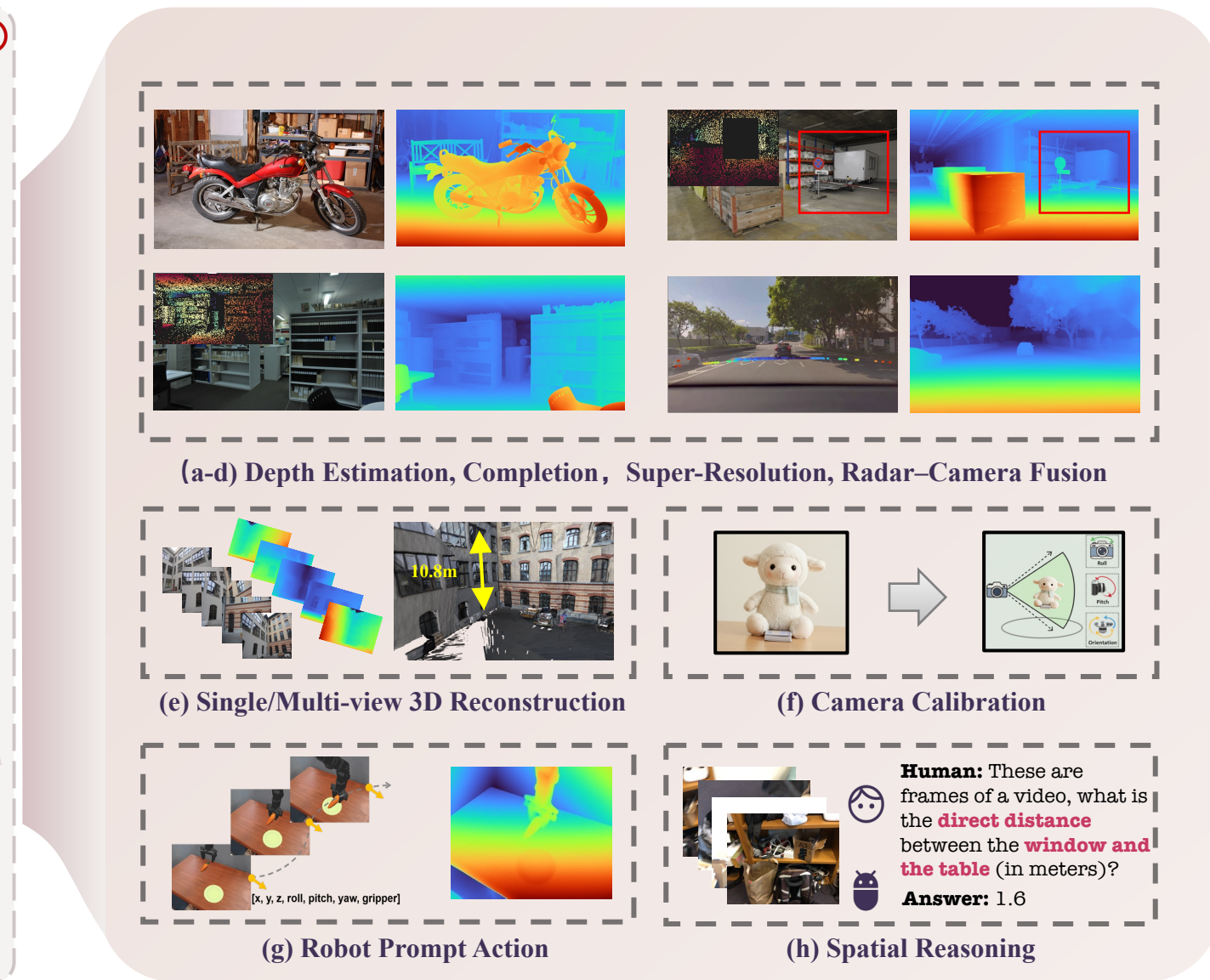**Answer:** 1.6

I. Cross-Source 3D Data　　　　　　**II. Prompt-based Pretraining**　　　　　　**III. Downstream Task**