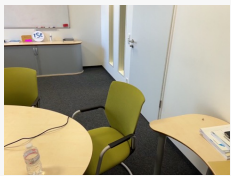


Input Image (I)



unify focal length

Text Prompt (T)

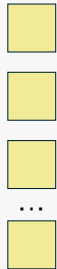
Given this image,
estimate the metric
depth (in meters)
for every pixel of
the image.

Multi-Layer Visual Features in ViT

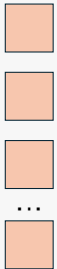


Vision
Encoder

Input Tokens



Generated Tokens

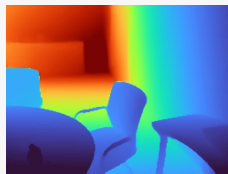


Large Language Model

Depth Prediction
Head

Text Decoding

Training Stage 1
+ Training Stage 2

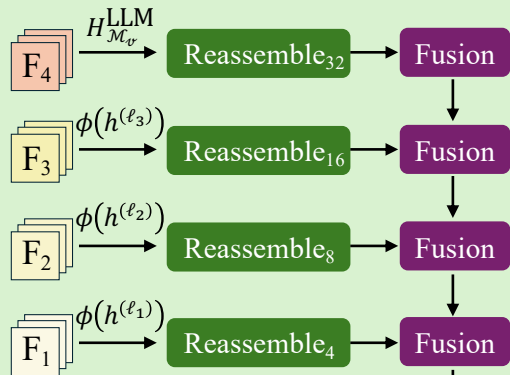


Metric Depth Map ($\hat{D}$)

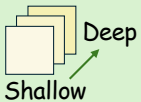
Text Response ($\hat{T}$)

A dense per-pixel
metric depth map
(in meters) has
been generated for
the input image.

Depth Prediction Head



Multi-Layer
Features in ViT



LLM Output
Visual Feature

