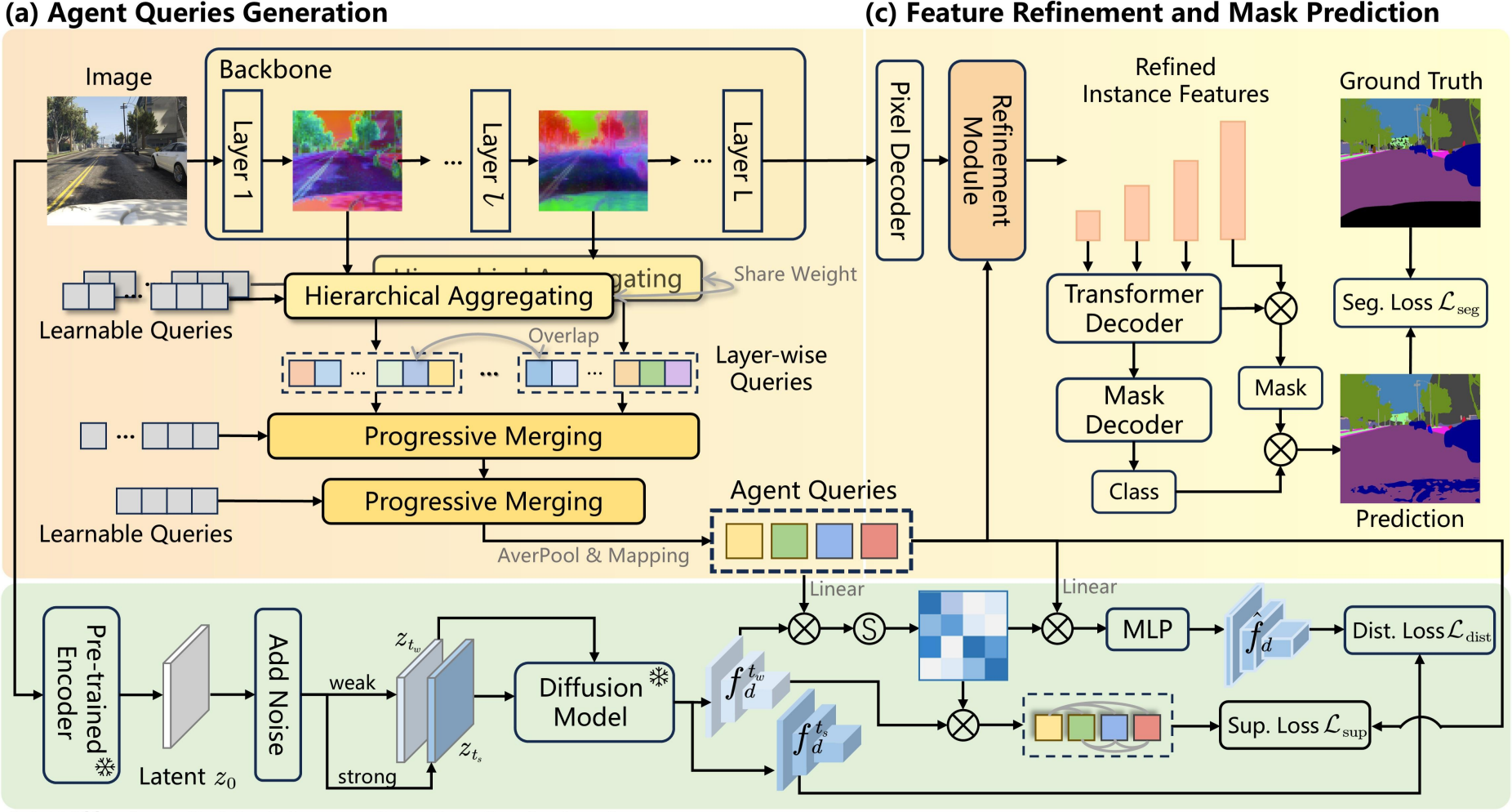
(a) Agent Queries Generation
(c) Feature Refinement and Mask Prediction
Image
Backbone
Layer 1
Layer l
Layer L
Pixel Decoder
Refinement Module
Refined Instance Features
Ground Truth
Share Weight
Hierarchical Aggregating
Learnable Queries
Overlap
Layer-wise Queries
Transformer Decoder
Seg. Loss $\mathcal{L}_{seg}$
Progressive Merging
Mask Decoder
Mask
Progressive Merging
Class
Agent Queries
Learnable Queries
AverPool & Mapping
Prediction
Linear
Linear
Pre-trained Encoder
Latent z_0
Add Noise
weak
strong
z_{t_w}
z_{t_s}
Diffusion Model
MLP
Dist. Loss $\mathcal{L}_{dist}$
Sup. Loss $\mathcal{L}_{sup}$
(b) Diffusion-Guided Agent Queries Optimization