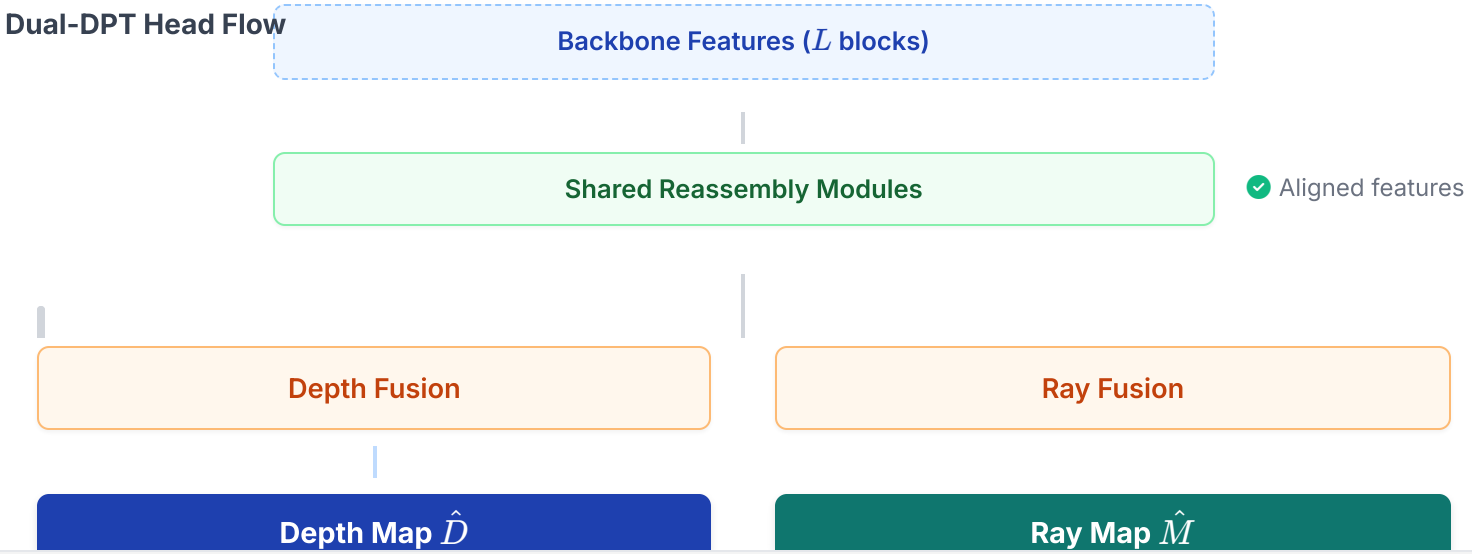
6 Architecture II: Dual-DPT Head & Camera Head



Camera Head D_C

Input: Camera Token

One token per view (c_i)

Predicts: f, q, t

Multi-Task Efficiency

Dual-DPT outperforms separate heads in pose and geometry (Tables 6-7).

Dual-DPT Head Design

- **Shared Reassembly:** Upsamples and concatenates multi-scale ViT features (from different transformer layers) into dense spatial representations before task-specific fusion (from DPT architecture).
- **Branch-Specific Fusion:** Two distinct paths fuse features for depth vs. ray tasks.
- **Benefit:** Encourages strong task interaction while minimizing redundant computation—outperforms separate heads with minimal parameter overhead.

Camera Head (D_C)

- Operates exclusively on **camera tokens** (one per view).
- Predicts explicit parameters: FOV $f \in \mathbb{R}^2$, quaternion $q \in \mathbb{R}^4$, translation $t \in \mathbb{R}^3$.
- **Efficiency:** Negligible computational cost—amortizes pose extraction without expensive dense ray map processing at inference.