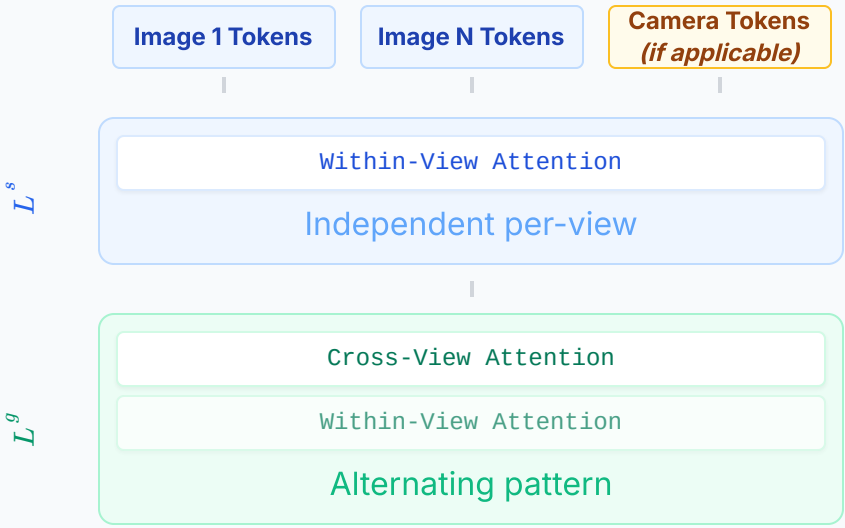
5 Architecture I: Single ViT & Input-Adaptive Attention

Transformer Block Structure

Pretrained DINOv2



KEY INNOVATION: Input-Adaptive Cross-View Attention

Standard ViT handles multiple views without modification

Phase 1: Within-View (L_s layers)

- Tokens attend only within their own view
- Extract per-view features (monocular depth)
- Build local context first

Phase 2: Cross-View (L_g layers)

Alternating:

- **Cross-view:** Reorder → interleave → correspondences
- **Within-view:** Group back → refine

Standard ViT: only token ordering changes

Input-Adaptive Property

- $N_v = 1$: Monocular depth (zero overhead)
- $N_v = 2$-18: Scales gracefully
- **Complexity:** $O(N^2)$ ViT (no 3D volumes)

Empirical optimal ratio: $L_s : L_g = 2 : 1$

Why 2:1 ratio?

Balances local feature learning with cross-view reasoning.

Trade-offs:

- Too few L_g → limited multi-view fusion
- Too many L_g → reduces within-view discriminability
- L_g scales as $O(N_v \cdot H \cdot W)^2$ vs $O(H \cdot W)^2$ for L_s

Tab. 7: 2:1 provides optimal accuracy-efficiency balance