Single-view transformer (SVT): process views separately
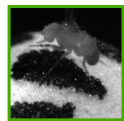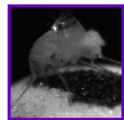
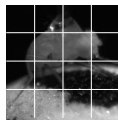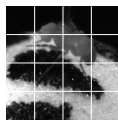Input images | Patchify | Patch Embed | Fixed Pos Enc | ViT | Heatmap Head | Predicted heatmaps (per keypoint) | Heatmap loss

View 1

View 2

Multi-view transformer (MVT): process views simultaneously
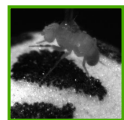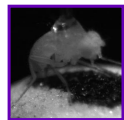
Input images | Patchify + Mask | Patch Embed | Fixed Pos Enc | Learnable View Enc | ViT | Heatmap Head | Predicted heatmaps (per keypoint) | Heatmap loss

View 1
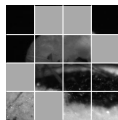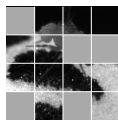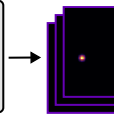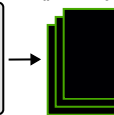
View 2

Predicted 3D keypoints → 3D keypoint loss

Legend:
- ■ Visible patch embedding
- ■ Masked patch embedding
- ◯ Position encoding
- ▽ View encoding