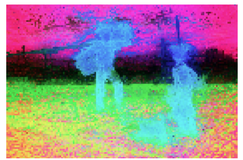
PCA maps: SigLIP2



DINOv3

Ours

Early-fusion Grounding VLM

Bounding box

Segmentation mask

SigLIP2-specific head

DINOv3-specific head

$$\mathcal{L} = \mathcal{L}^{(S)}_{\text{ARKD}} + \mathcal{L}^{(S)}_{\text{CLS}} + \mathcal{L}^{(S)}_{\text{patch}} + \mathcal{L}^{(D)}_{\text{ARKD}} + \mathcal{L}^{(D)}_{\text{CLS}} + \mathcal{L}^{(D)}_{\text{reg}} + \mathcal{L}^{(D)}_{\text{patch}}$$

SigLIP2

DINOv3

Vision Expert 1

Vision Expert 2

Text Expert 1

Text Expert 2

Shared Expert 1

MoE layer

Vision Expert 1

Vision Expert 2

Text Expert 1

Text Expert 2

Shared Expert 1

MoE layer

Text token

Register token

Global representation token

Patch token

Stage 0: MT-Distillation

Patch embedding layer

Embedding layer

The statue of the person holding two baskets