A General Protocol to Probe Large Vision Models for 3D Physical Understanding

Guanqi Zhan, Chuanxia Zheng, Weidi Xie, Andrew Zisserman

Visual Geometry Group, University of Oxford

NeurIPS 2024

Problem, Overview and Related Work

- **Problem**: To what extent these large-scale vision models (such as CLIP/DINO/Stable Diffusion/VQGAN) have learned about the 3D scene depicted with only the 2D images.

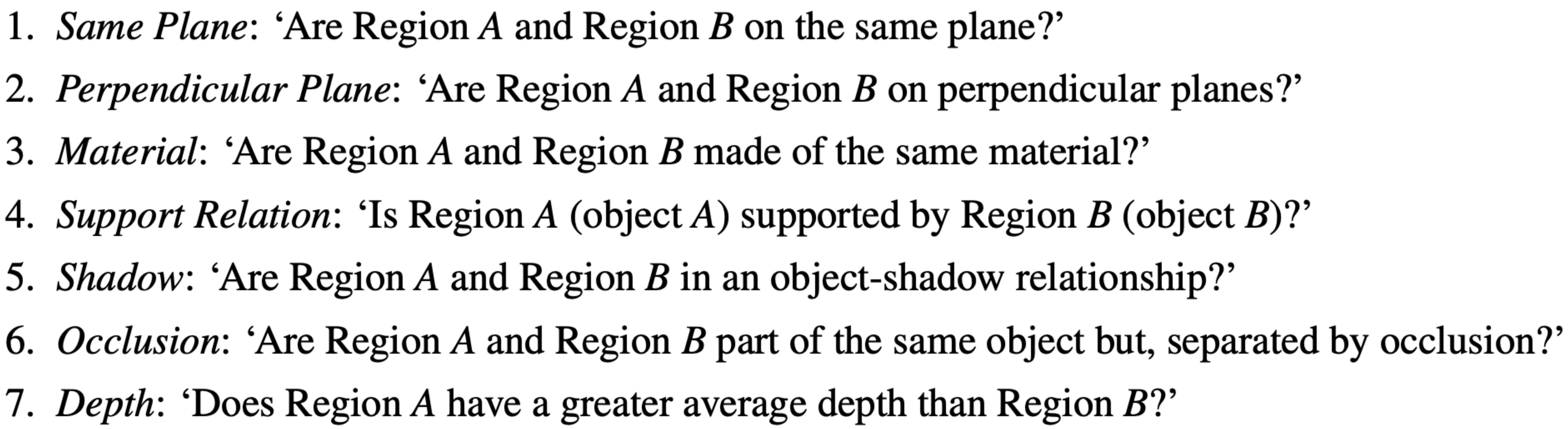
- **Overview**: A general protocol to probe the large-scale vision models
  - Set up the binary probe **questions** for each 3D physical property
  - Train a binary classifier **to probe** the features of large-scale vision models, and this indicates whether there are explicit features for such 3D scene properties

- **Related Work**
  - Ayush Sarkar, Hanlin Mai, Amitabh Mahapatra, Svetlana Lazebnik, David Forsyth, Anand Bhattad. Shadows don’t lie and lines can’t bend! generative models don’t know projective geometry... for now. CVPR, 2024.
  - Mohamed El Banani, Amit Raj, Kevis-Kokitsi Maninis, Abhishek Kar, Yuanzhen Li, Michael Rubinstein, Deqing Sun, Leonidas Guibas, Justin Johnson, Varun Jampani. Probing the 3D awareness of visual foundation models. CVPR, 2024.

Protocol: Properties and Questions

- **Properties**: scene geometry, material, support relations, shadows, occlusion and depth.

- **Questions**: For each property, we propose questions that classify the relationship between a pair of Regions, A and B, in the same image, based on the features extracted from the large vision model



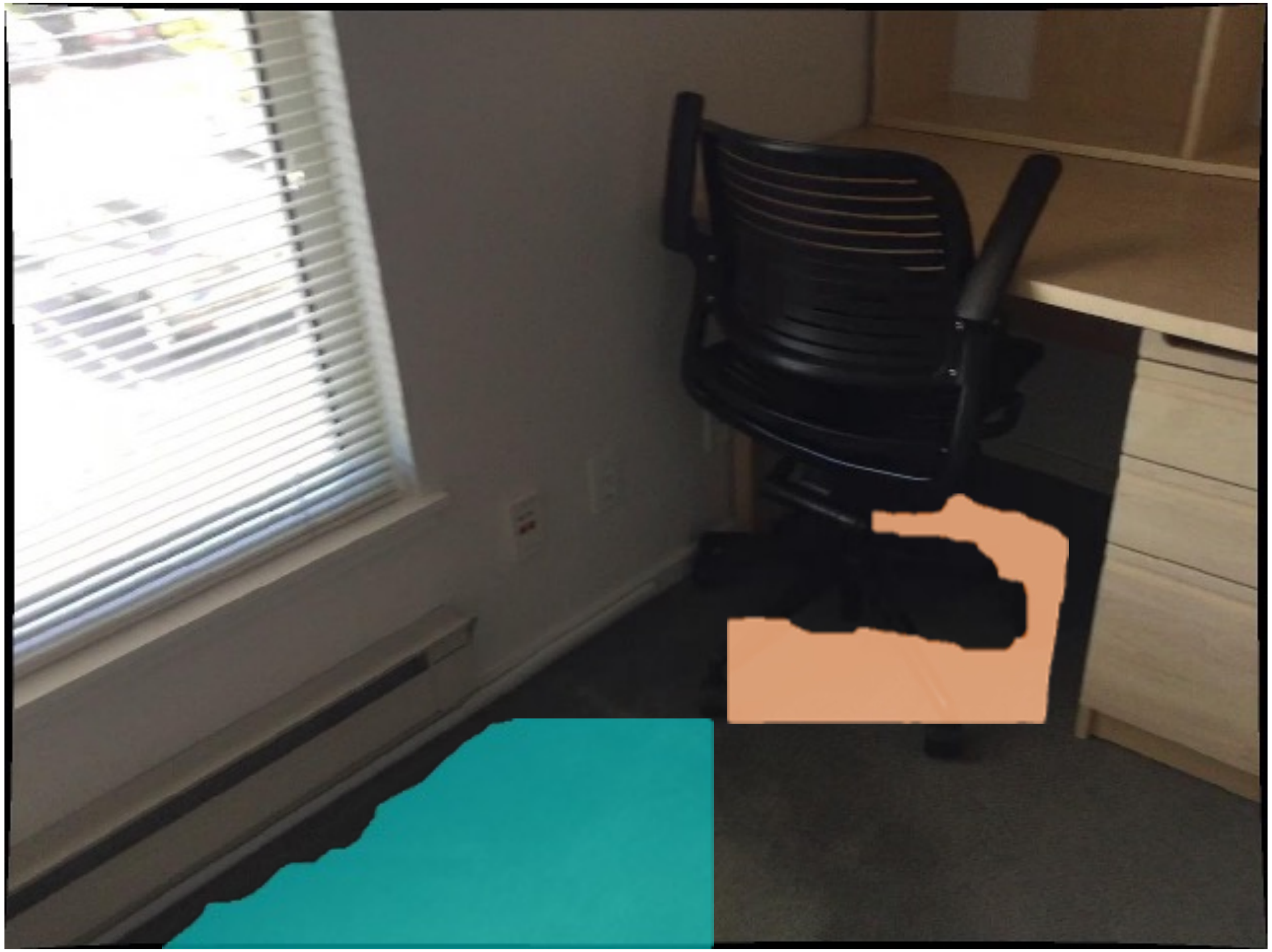
1. *Same Plane*: ‘Are Region *A* and Region *B* on the same plane?’
2. *Perpendicular Plane*: ‘Are Region *A* and Region *B* on perpendicular planes?’
3. *Material*: ‘Are Region *A* and Region *B* made of the same material?’
4. *Support Relation*: ‘Is Region *A* (object *A*) supported by Region *B* (object *B*)?’
5. *Shadow*: ‘Are Region *A* and Region *B* in an object-shadow relationship?’
6. *Occlusion*: ‘Are Region *A* and Region *B* part of the same object but, separated by occlusion?’
7. *Depth*: ‘Does Region *A* have a greater average depth than Region *B*?’

Examples of Protocol for Same Plane

- **Are Region A and Region B on the same plane?**

Examples of Protocol for Same Plane

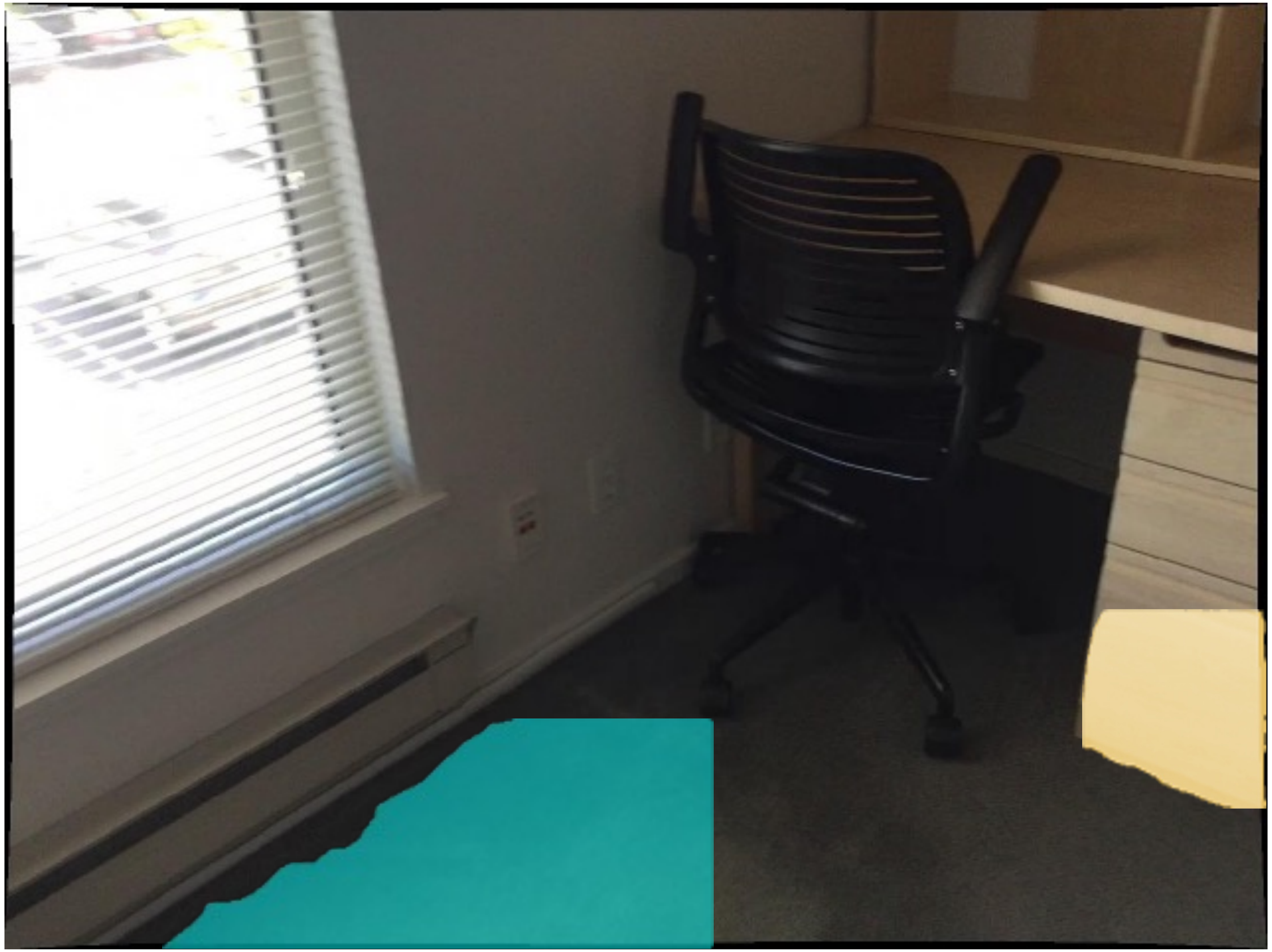
- **Are Region A and Region B on the same plane?**



Yes

Examples of Protocol for Same Plane

- **Are Region A and Region B on the same plane?**



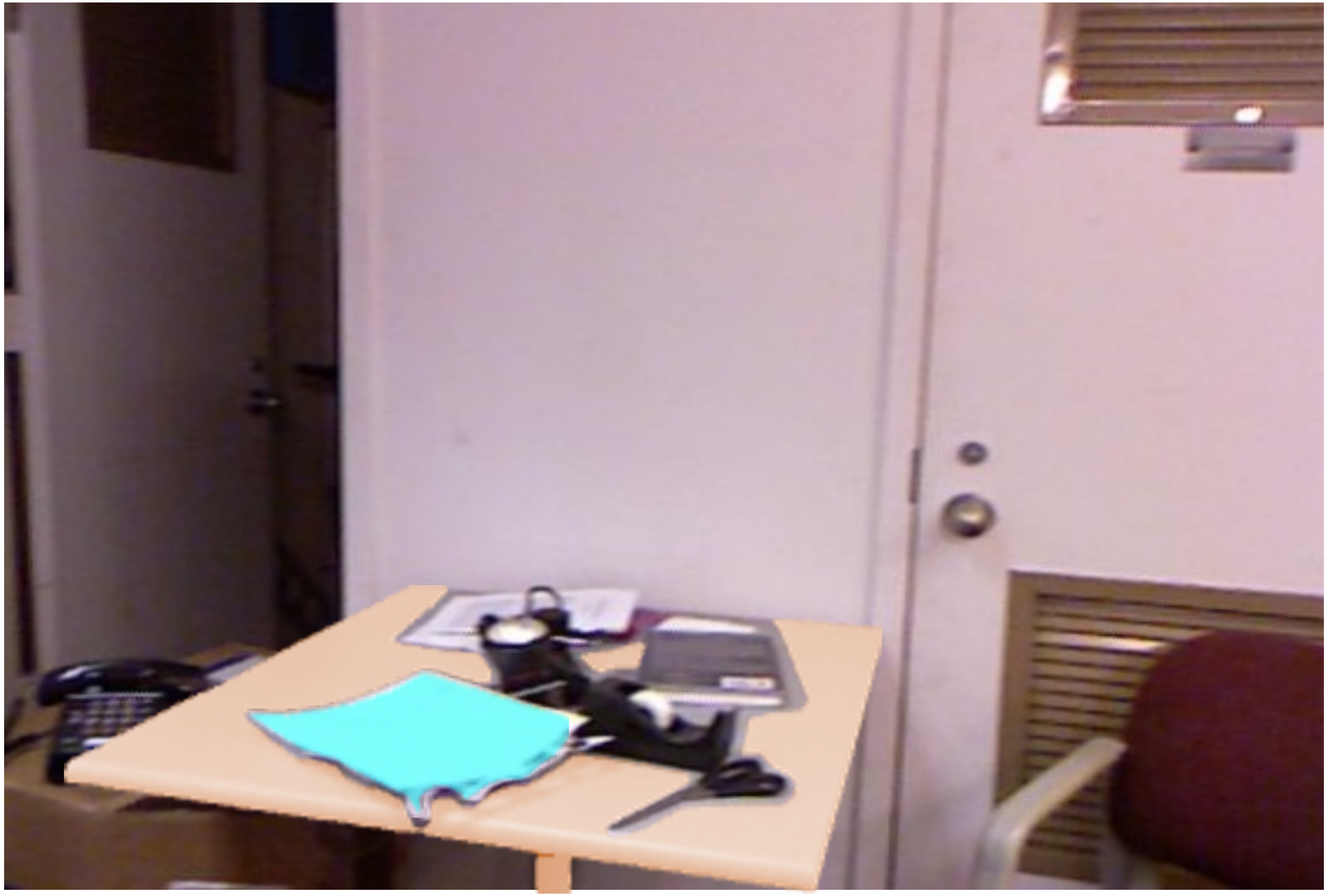
No

Examples of Protocol for Support Relation

- **Is Region A supported by Region B?**

Examples of Protocol for Support Relation

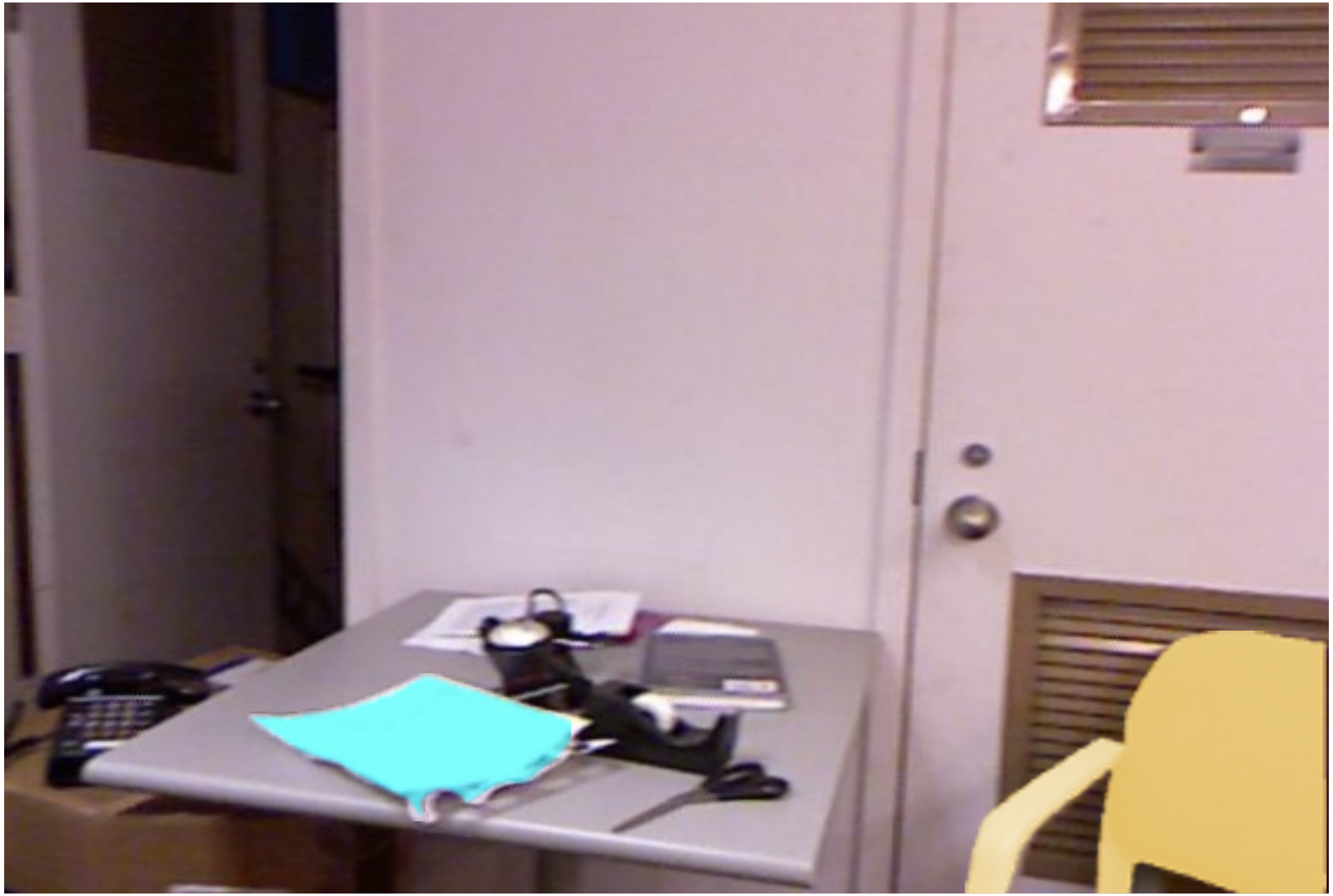
- **Is Region A supported by Region B?**



Yes

Examples of Protocol for Support Relation

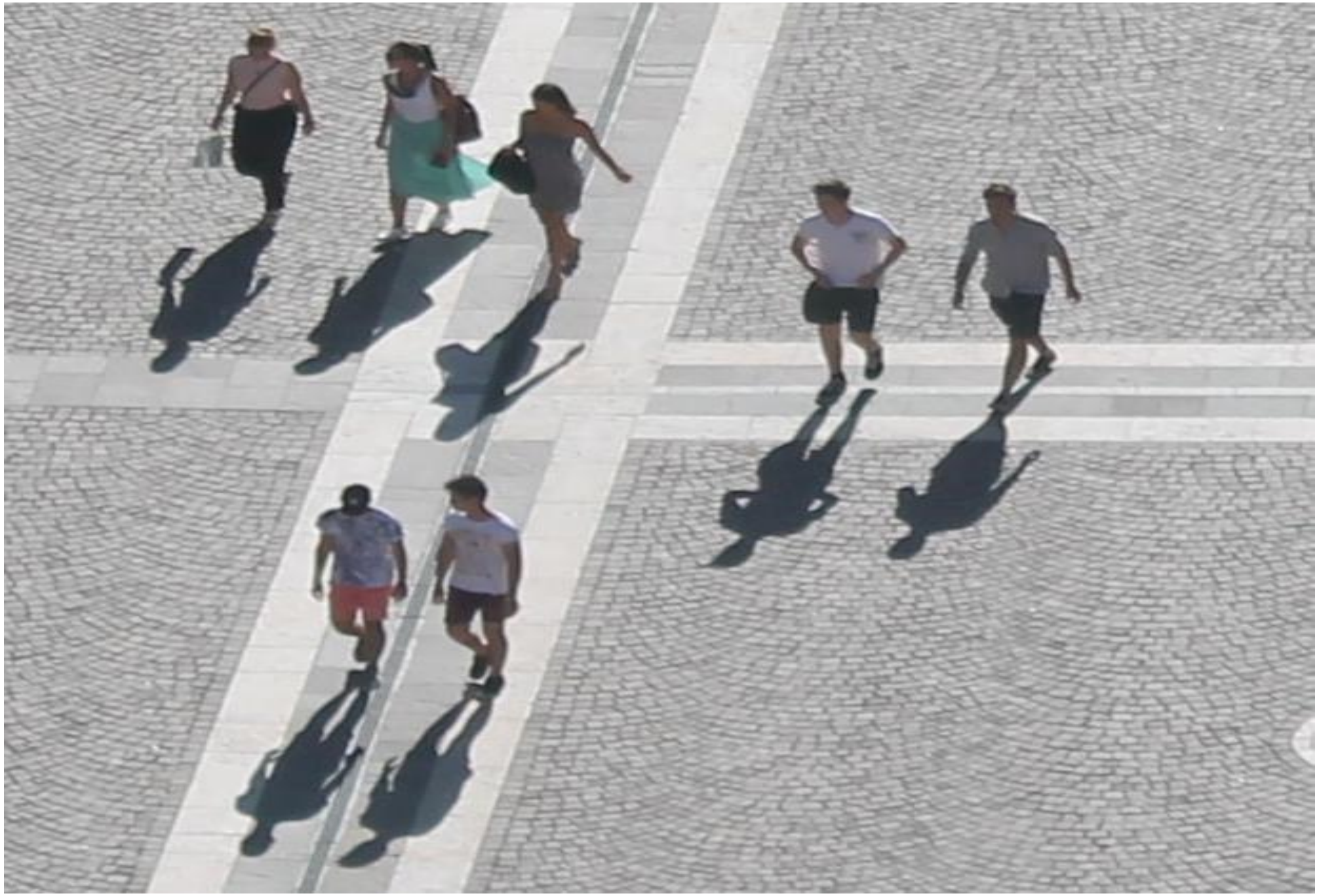
- **Is Region A supported by Region B?**



No

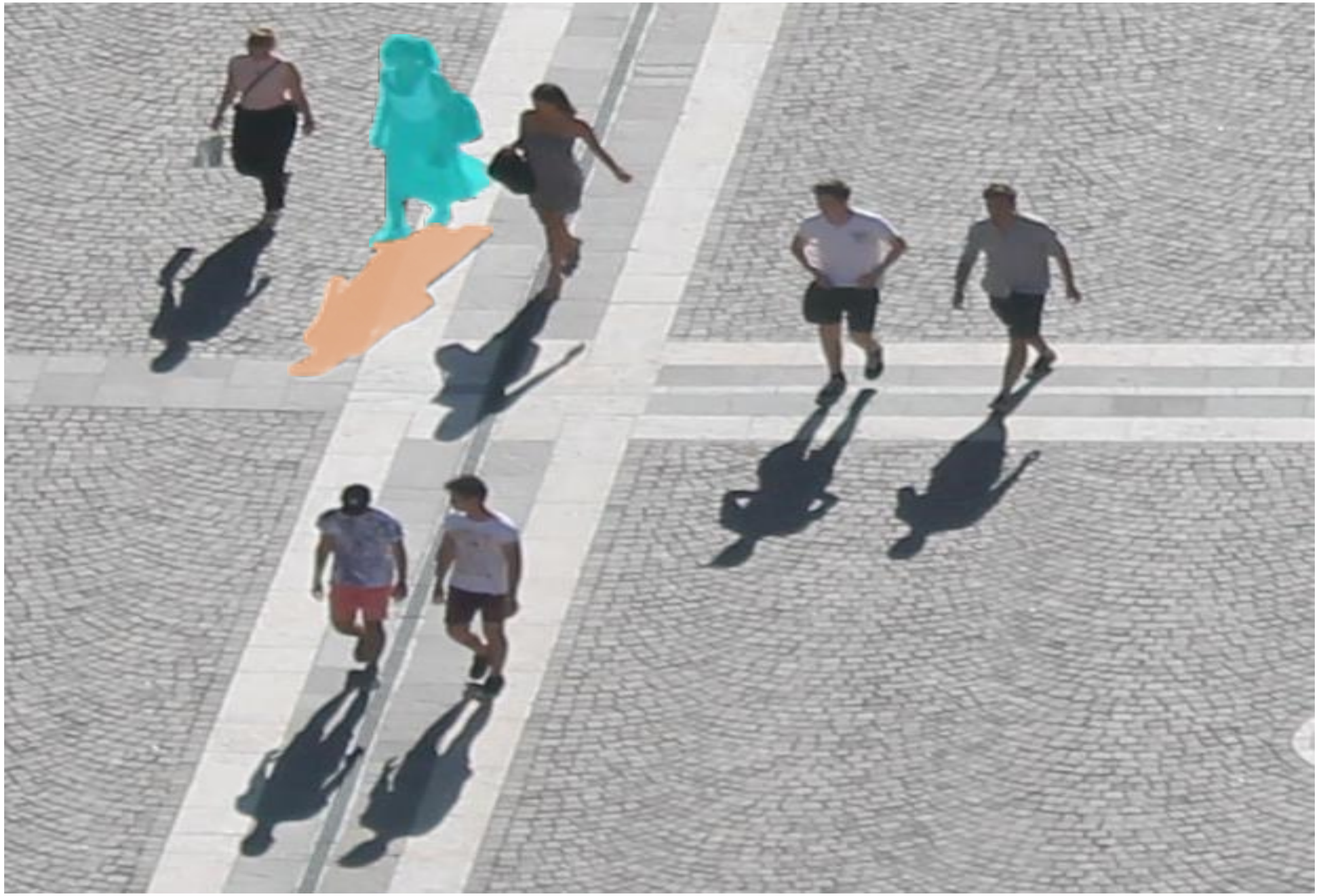
Examples of Protocol for Shadow

- **Are Region A and Region B in an object-shadow relationship?**



Examples of Protocol for Shadow

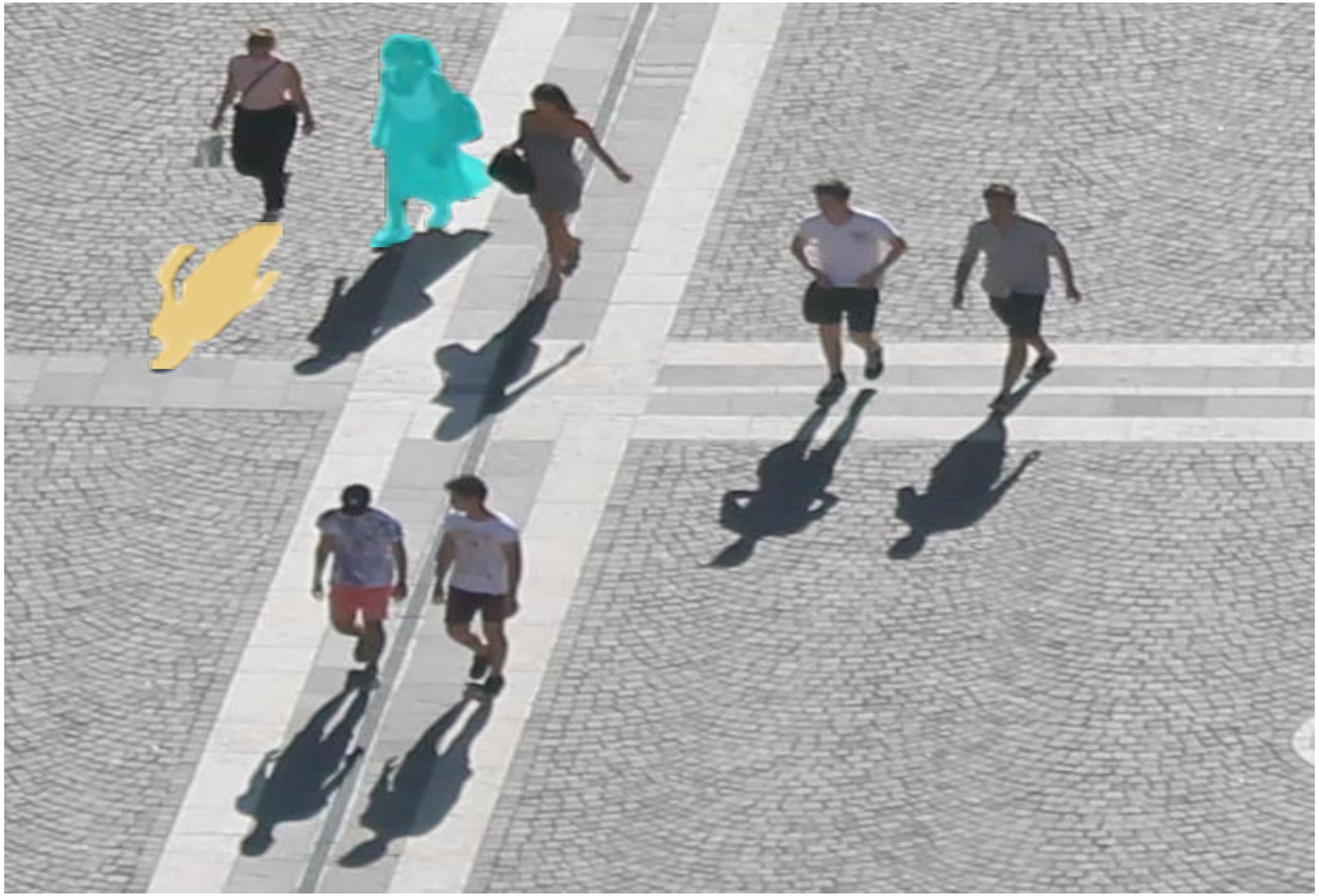
- **Are Region A and Region B in an object-shadow relationship?**



Yes

Examples of Protocol for Shadow

- **Are Region A and Region B in an object-shadow relationship?**



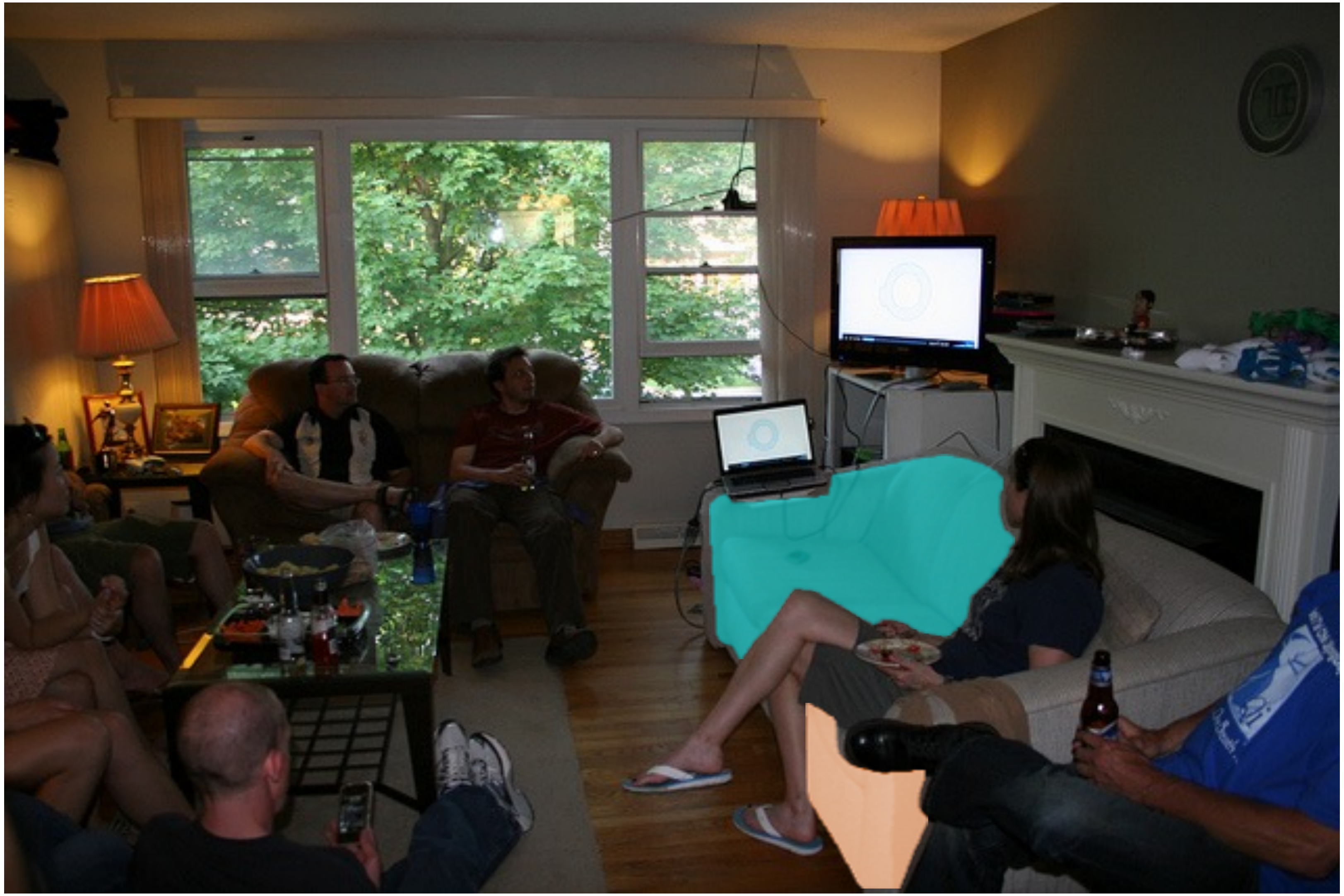
No

Examples of Protocol for Occlusion

- **Are Region A and Region B part of the same object but, separated by occlusion?**

Examples of Protocol for Occlusion

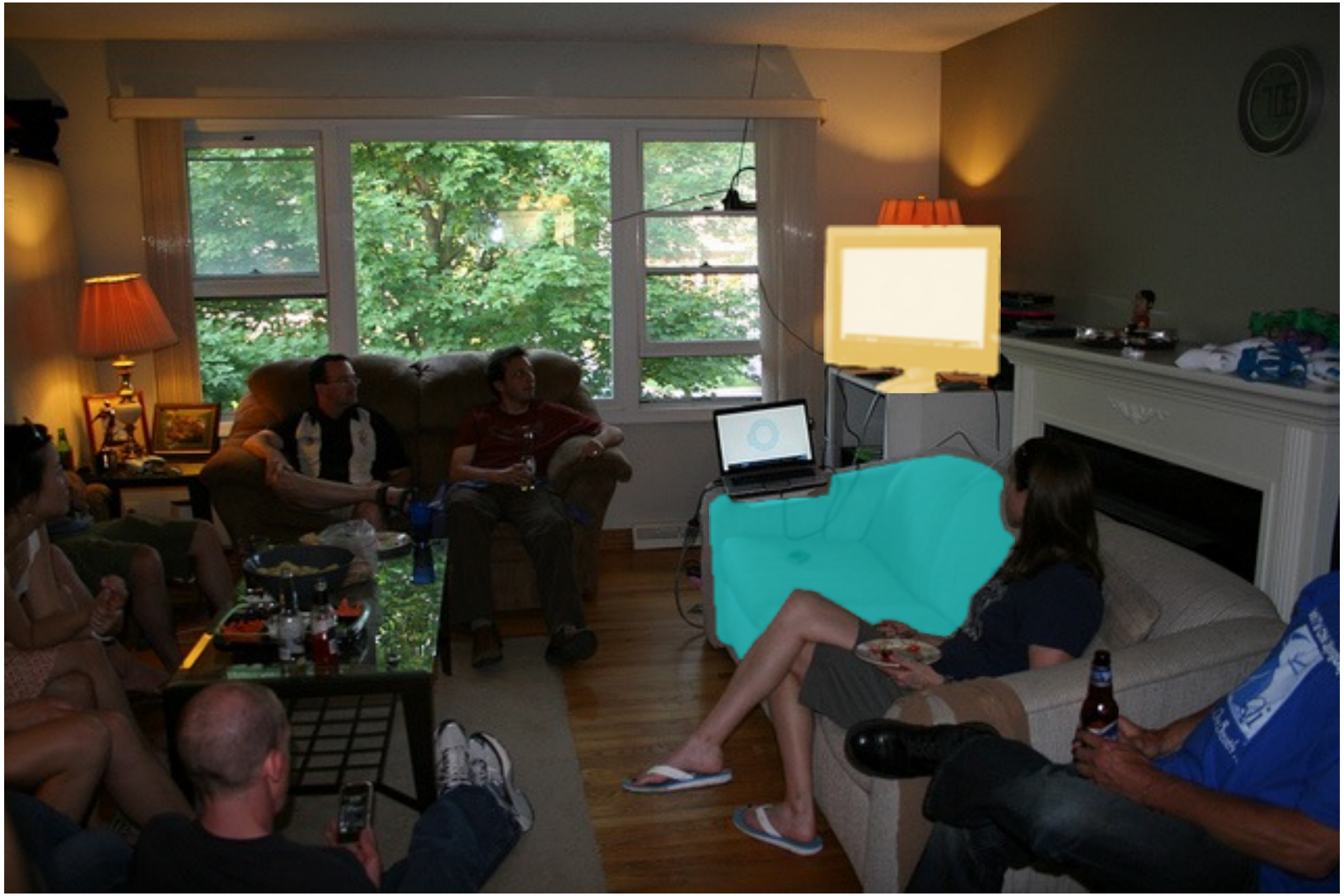
- **Are Region A and Region B part of the same object but, separated by occlusion?**



Yes

Examples of Protocol for Occlusion

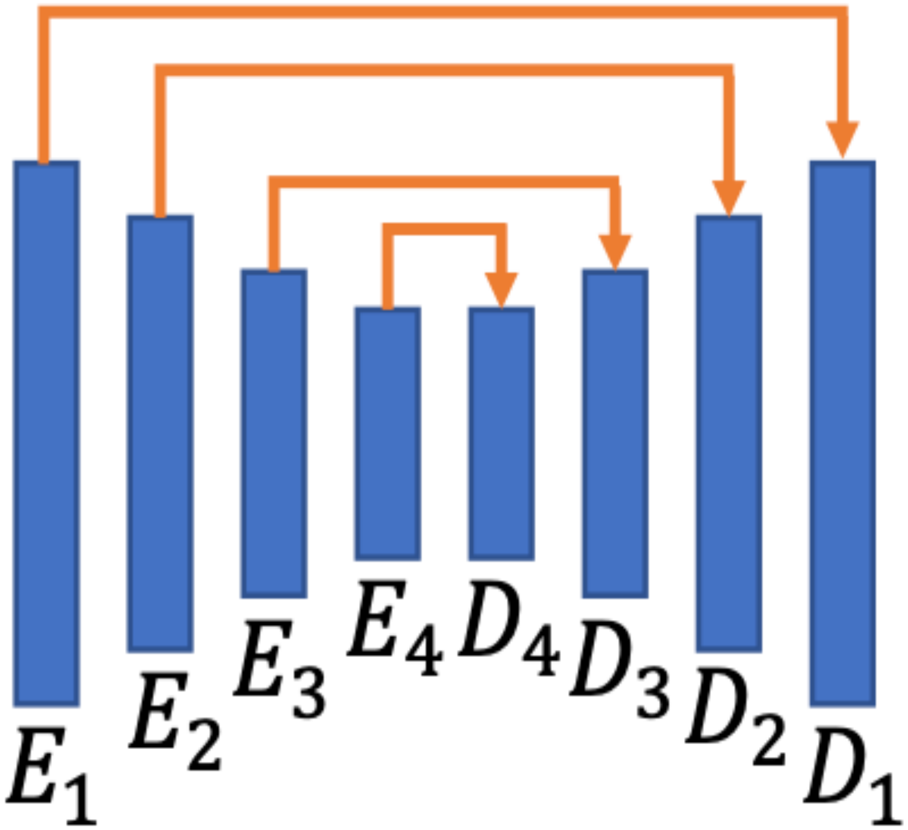
- **Are Region A and Region B part of the same object but, separated by occlusion?**



No

Protocol: Probe Stable Diffusion

- **Goal: Find the layer and time step which can best answer the questions**
  - **The way of doing it is to train a linear SVM as a binary classifier on features from different layers and time steps**



Encode image with VAE into the latent space z_0 and then add noise:

$$z_t = \sqrt{\alpha_t} z_0 + (\sqrt{1-\alpha_t})\epsilon$$

Obtain Stable Diffusion feature at timestep t, layer l:

$$F_{t,l} = f_{\theta_l}(z_t, t)$$

Obtain feature for region k:

$$v_{k,t,l} = \text{avgpool}(R_k \odot \text{upsample}(F_{t,l}))$$

Binary classification via linear SVM:

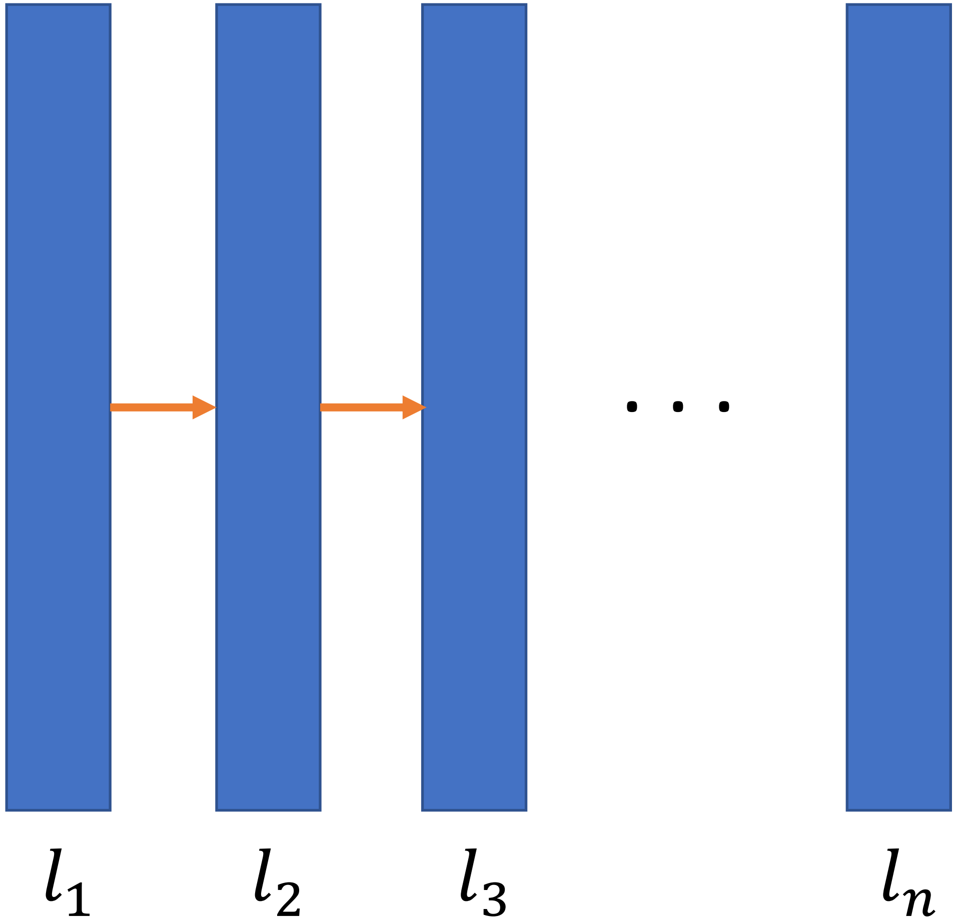
$$\text{SVM is given by } sign(w^T v + b)$$

Grid search

- (i) the optimal time step t
- (ii) the optimal U-Net layer l
- (iii) the SVM regularization parameter C

Protocol: Probe Probe CLIP/DINOv1/DINOv2/VQGAN

- **Goal: Find the layer which can best answer the questions**
  - **The way of doing it is to train a linear SVM as a binary classifier on features from different transformer layers**



Obtain CLIP/DINO/VQGAN feature at transformer layer l:

$$F_l = f_{\theta_l}(F_{l-1})$$

Obtain feature for region k:

$$v_{k,l} = \text{avgpool}(R_k \odot \text{upsample}(F_l))$$

Binary classification via linear SVM:

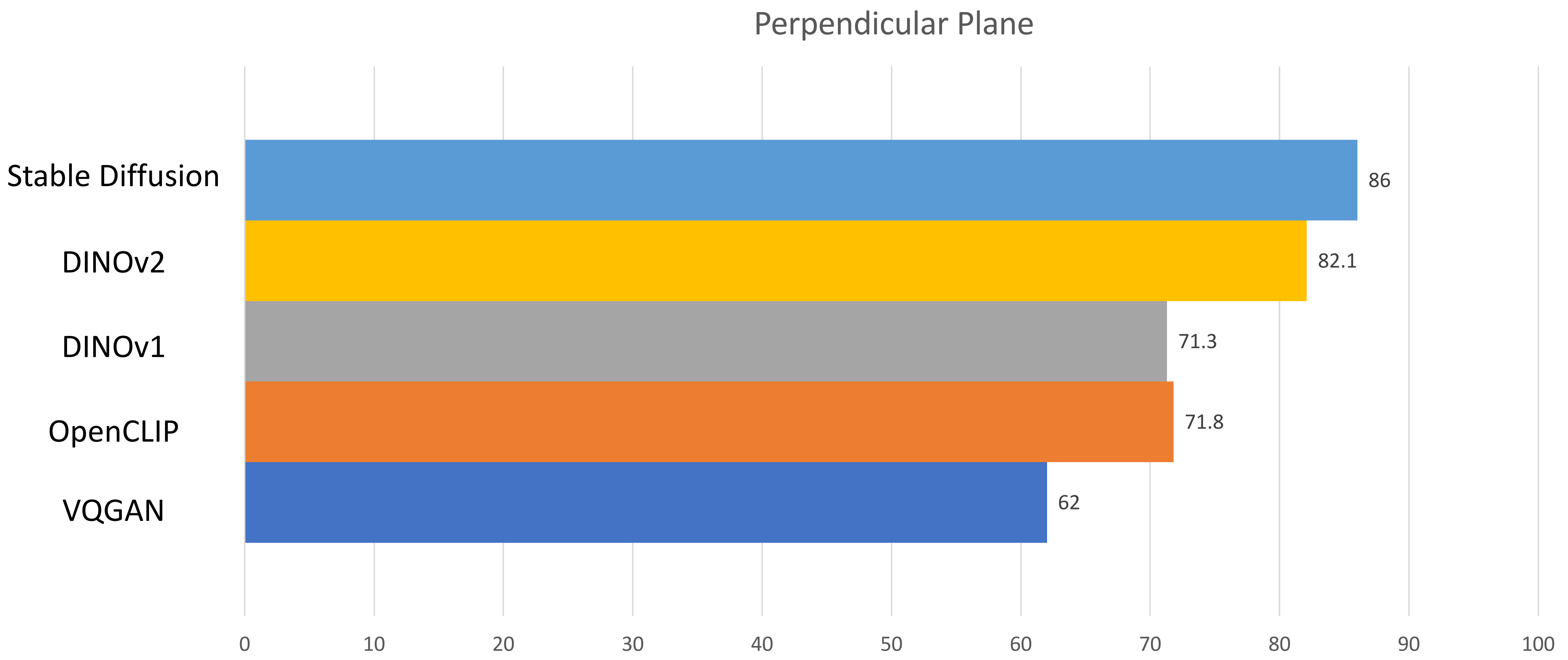
$$\text{SVM is given by } sign(w^T v + b)$$

Grid search

- (i) the optimal Transformer layer l
- (ii) the SVM regularization parameter C

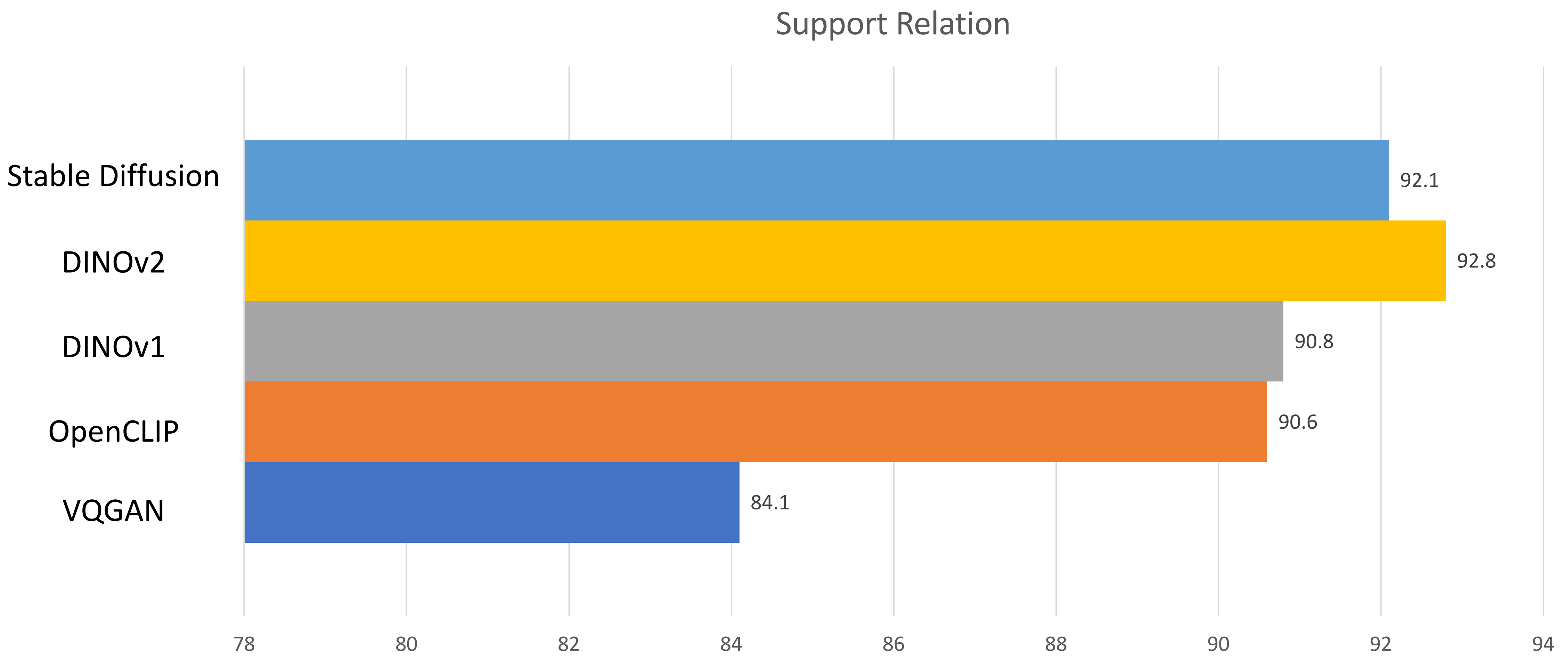
Results – Perpendicular Plane

- **Comparison of different features trained at scale**



Results – Support Relation

- **Comparison of different features trained at scale**



Conclusion

- **Findings:**
  - For Stable Diffusion, the most efficient feature is usually at the decoder side, and the best time step is usually before 400 as there will be too much noise after 400 time steps.
  - For CLIP/DINO/VQGAN, different layers of different models are good at different properties.
  - Features from Stable Diffusion and DINOv2 are good for discriminative learning of a number of properties, including scene geometry, support relations, shadows and depth, but less performant for occlusion and material.
  - Stable Diffusion and DINOv2 outperform DINOv1, CLIP and VQGAN for all properties.
- **Further Application:**
  - Potential usage of Stable Diffusion and DINOV2 features to enable 3D physical understanding tasks in the wild.
  - By knowing what properties Stable Diffusion is not good at, we have a way to spot images generated by Stable Diffusion.
  - It also reveals which properties the network could be trained further on to improve its 3D modelling, e.g., via extra supervision for that property.

Datasets and code are available

NEURAL INFORMATION PROCESSING SYSTEMS

VGG
UNIVERSITY OF OXFORD

UNIVERSITY OF
OXFORD

Thank you!