NEURAL INFORMATION PROCESSING SYSTEMS

NeurIPS 2026 Workshop
TAE (Trust-AI-Eval): Can We Trust AI Evaluation?

UNIVERSITÄT MANNHEIM

MAX PLANCK INSTITUTE FOR INFORMATICS

When Evaluation Goes Blind

Analyzing Patch-Grid Failure Modes in Vision Transformers

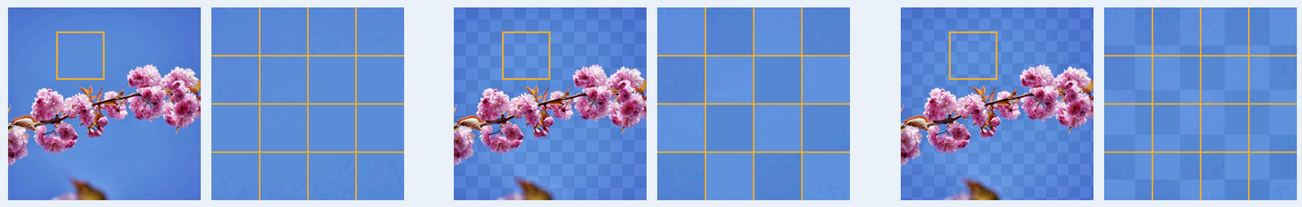
Steffen Jung · Margret Keuper
University of Mannheim · Max Planck Institute for Informatics, Saarland Informatics Campus

vitgrid.jung.vision

KEY FINDING

Same strength, one pixel apart: a checkerboard that fits the ViT patch grid is almost invisible to the encoder.

d_{rel}: change of the CLIP ViT-L/14 embedding, relative to the mean distance between two clean images. Amber: the 14-px patch grid. Close-ups show the boxed region at true strength (a = 0.025).

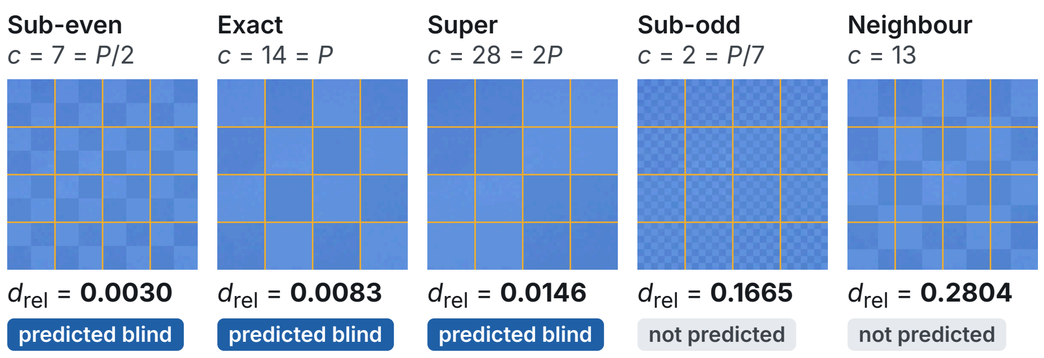


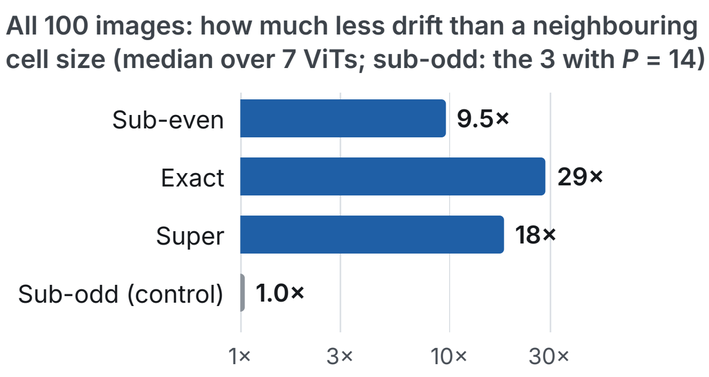
Setup. Learned evaluators (CLIPScore, PickScore, ImageReward, VQAScore) score images with ViT encoders that cut images into non-overlapping $P \times P$ patches. We add a checkerboard $\delta(u,v) = a \cdot (-1)^{\lfloor (u+\varphi_x)/c \rfloor + \lfloor (v+\varphi_y)/c \rfloor}$ with cell width c, offset φ and matched RMS strength a to 100 DIV2K images, then compare a predicted-blind cell width with its neighbour: **× less drift** = drift(neighbour) ÷ drift(predicted blind), median over images.

VARY: CELL WIDTH c

1 Which cell sizes are blind?

Same image and strength, zero offset; P = 14. The patch-grid rule predicts **sub-even**, **exact** and **super** cells to be blind; **sub-odd** cells are a control.



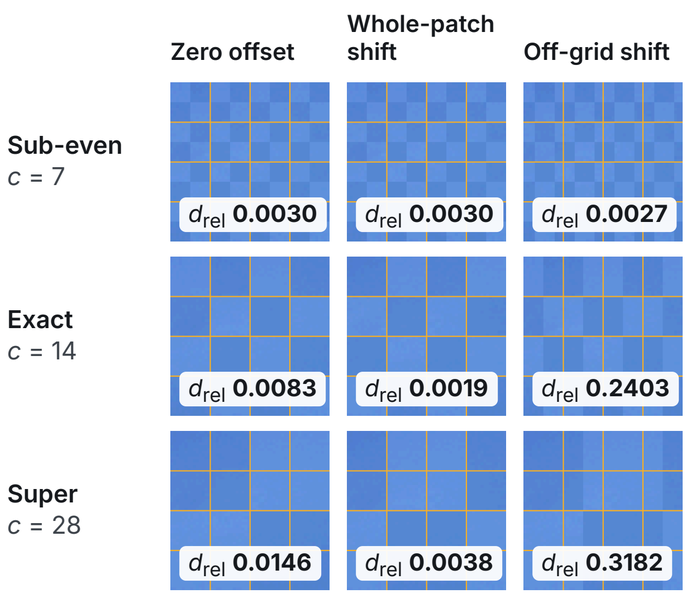


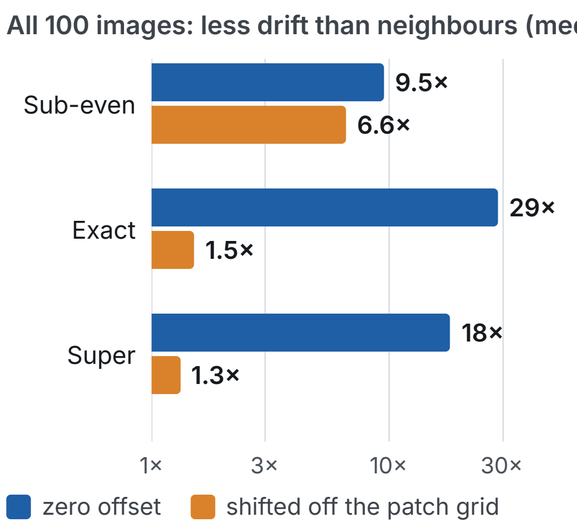
VARY: OFFSET φ

2 Size or alignment?

A shift keeps the frequency but moves cell edges relative to the fixed patch grid.

Off the grid, exact and super cells lose their blind spot; sub-even cells stay blind. Frequency alone does not explain the effect.

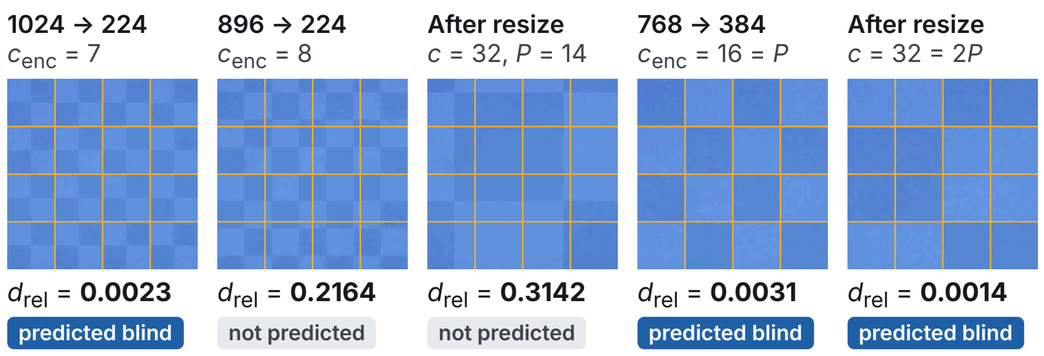




VARY: WHERE THE PATTERN IS ADDED

3 Resizing moves the blind spot

The same 32-px cell becomes c_{enc} = 7 px after 1024→224 resizing (sub-even) but 8 px after 896→224. Drift follows the cell width the encoder sees.



All 100 images: when resizing changes which setting is predicted blind, that setting has lower median drift in

84/90 comparisons, five 224-px encoders

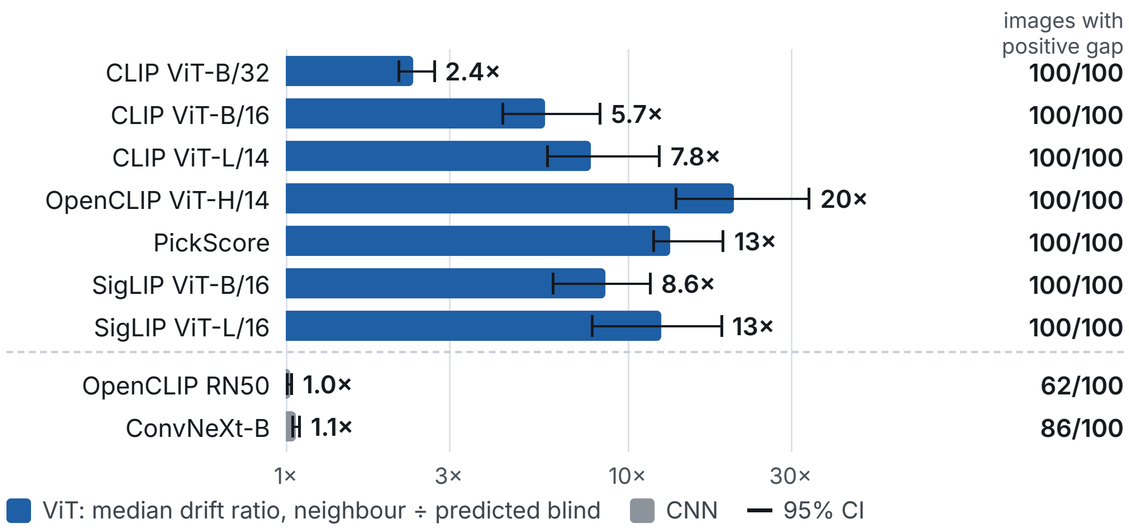
6/6 comparisons, both 384-px SigLIP models

The two 384-px tiles use the SigLIP grid (P = 16) and SigLIP ViT-B/16 drift; the others use CLIP ViT-L/14.

VARY: ENCODER

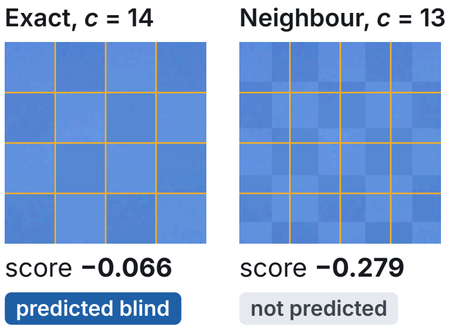
4 Every ViT is affected — CNNs barely

All eligible cell-size pairs and strengths. ViTs patchify in one step; CNNs downsample in several (P = 32 = their stride).



DOWNSTREAM: LAION-AESTHETICS ON CLIP VIT-L/14

5 The blind spot reaches the score



Clean score 5.255. One pixel of cell width makes the score drop 4.2× more for this image.

4.9× smaller median score change: 0.026 vs 0.128; gap positive in 92/100 images

6% vs 24% of 1,700 image pairs flip a keep/reject decision at a score cutoff of 5.5

Take-away. ViT-based evaluators have reproducible, patch-grid-shaped blind spots. Test them with structured perturbations at several sizes and offsets before trusting their scores.

Supported by BMFTR project TrackOpt (01IS24074A-D) and DFG Research Unit 5336 – Learning to Sense.

INTERACTIVE PROJECT PAGE
vitgrid.jung.vision

steffen.jung@uni-mannheim.de