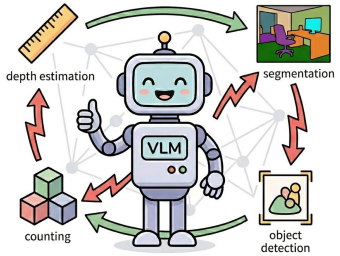
Microsoft Research



CVPR JUNE 3-7, 2026 DENVER COLORADO

Understanding Task Transfer in Vision-Language Models

*Bhuvan Sachdeva**, *Karan Uppal**, *Abhinav Java**, *Vineeth N. B.*

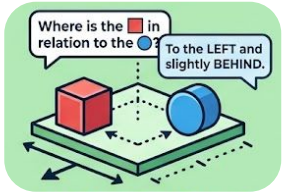
Microsoft Research India

https://aka.ms/task-transfer-vlms

Finetuning VLMs on Perception Tasks

Perception tasks humans find natural are quite hard for current VLMs

Counting 🔢



Spatial Reasoning 🏔️

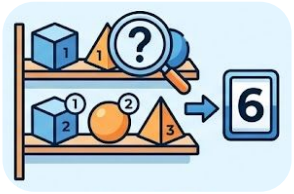
Relative Depth 📏

Object Localization 🎯

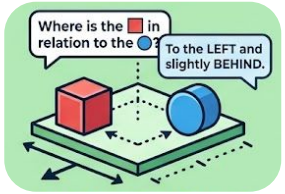
....

Finetuning VLMs on Perception Tasks

Perception tasks humans find natural are quite hard for current VLMs



Counting 🔢



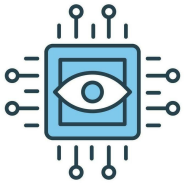
Spatial Reasoning 🏔️

Relative Depth 📏

Object Localization 🎯

....

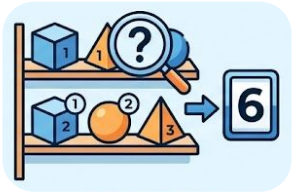
Current Scenario



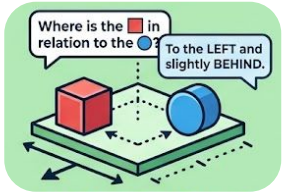
VLM → Finetune on one task 📐 → Good results on 📐⬆

Finetuning VLMs on Perception Tasks

Perception tasks humans find natural are quite hard for current VLMs



Counting 🔢



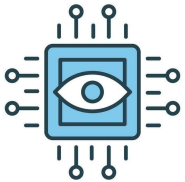
Spatial Reasoning 🏔️

Relative Depth 📏

Object Localization 🎯

....

Current Scenario



VLM → Finetune on one task 📐 → Good results on 📐⬆️

↓

Unknown interferences on other tasks 🔢 🏔️ 📏 🎯 🧩 🖼️

No understanding of how tasks interact with each other!

Prior Work

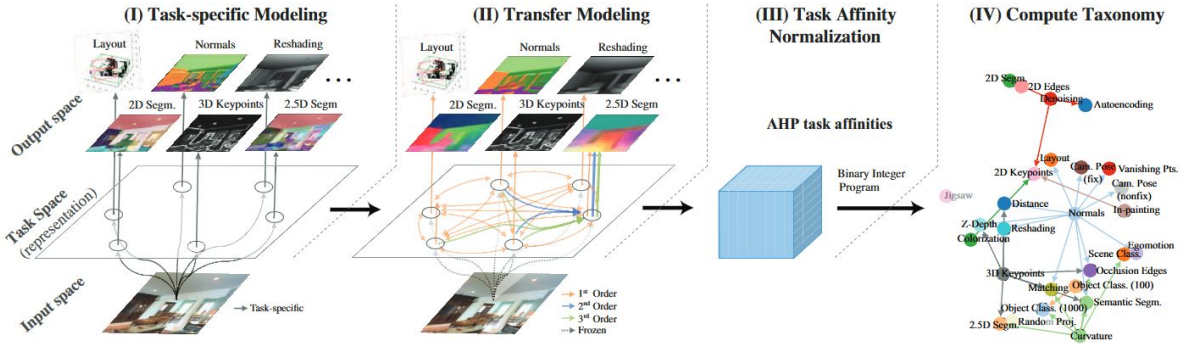
Taskonomy: Disentangling Task Transfer Learning

Amir R. Zamir[1,2] Alexander Sax[1*] William Shen[1*] Leonidas Guibas[1] Jitendra Malik[2] Silvio Savarese[1]

[1] Stanford University [2] University of California, Berkeley

http://taskonomy.vision/

Fun Fact: Taskonomy won the CVPR 2018 Best Paper Award



Prior Work

Pretraining focused transfer studies

How well do contrastively trained models transfer?

M. Moein Shariatnia [*1] Rahim Entezari [*2] Mitchell Wortsman [3] Olga Saukh [2] Ludwig Schmidt [3]

Learning More May Not Be Better: Knowledge Transferability in Vision and Language Tasks

Tianwei Chen[1], Noa Garcia[1], Mayu Otani[2], Chenhui Chu[3], Yuta Nakashima[1], Hajime Nagahara[1]
Osaka University[1], CyberAgent Inc.[2], Kyoto University[3]
{chentw@is., noagarcia@, n-yuta@, nagahara@}ids.osaka-u.ac.jp
otani_mayu@cyberagent.co.jp
chu@i.kyoto-u.ac.jp

Evaluation focused

What Are We Measuring When We Evaluate Large Vision-Language Models? An Analysis of Latent Factors and Biases

Anthony Meng Huat Tiong[1*], Junqi Zhao[1*], Boyang Li[1♠], Junnan Li,
Steven C.H. Hoi[2], and Caiming Xiong[3]
[1]Nanyang Technological University [2]Singapore Management University [3]Salesforce Research
{anthonym001, junqi.zhao, boyang.li}@ntu.edu.sg
{junnan4926, stevenhoi}@gmail.com cxiong@salesforce.com

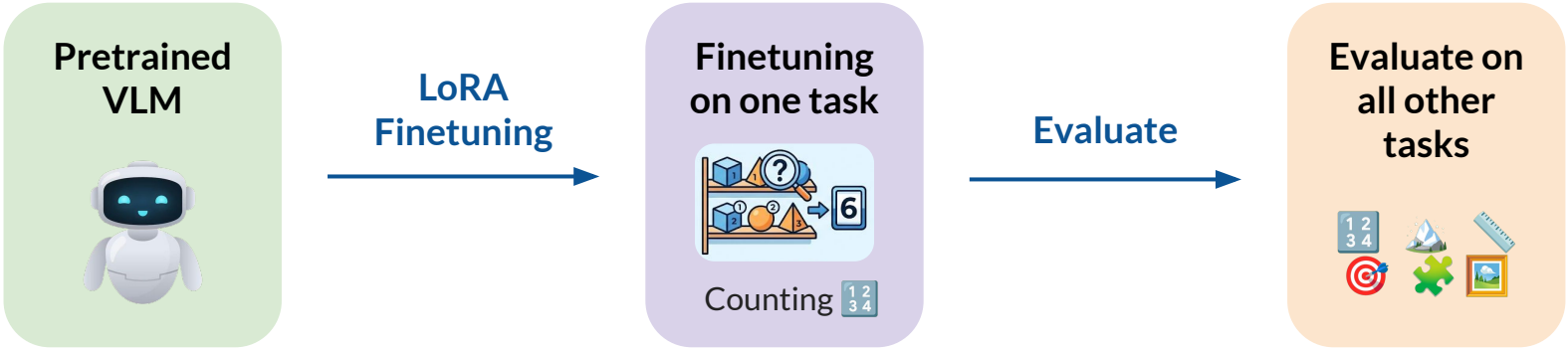
Effect of SFT vs RL in language

DOES MATH REASONING IMPROVE GENERAL LLM CAPABILITIES? UNDERSTANDING TRANSFERABILITY OF LLM REASONING

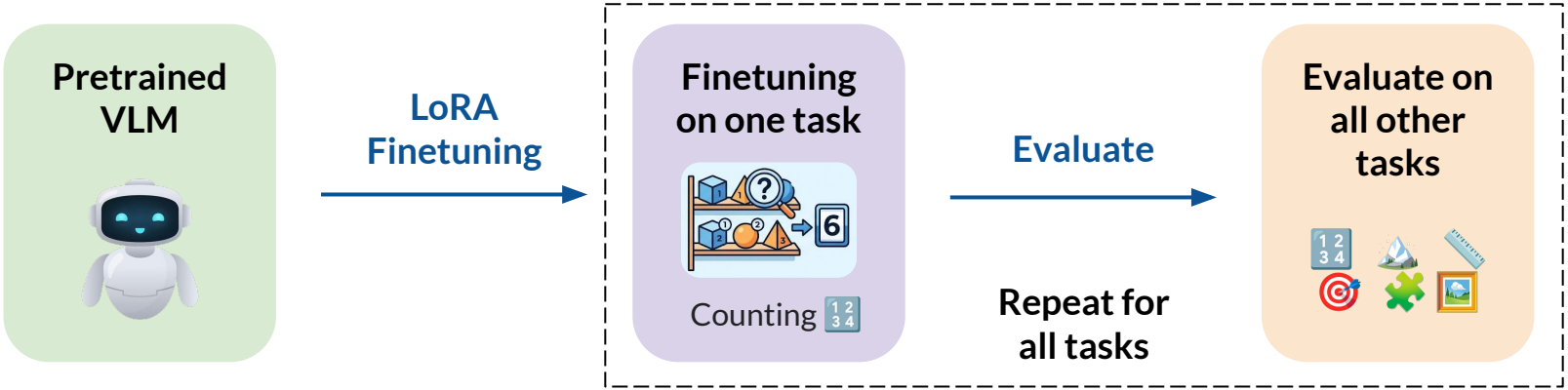
Maggie Huan[1,2*] Yuetai Li[3*] Tuney Zheng[4*] Xiaoyu Xu[5] Seungone Kim[1]
Minxin Du[5] Radha Poovendran[3] Graham Neubig[1] Xiang Yue[1†]
[1]Carnegie Mellon University [2]University of Pennsylvania [3]University of Washington
[4]M-A-P [5]The Hong Kong Polytechnic University
ziyuh@seas.upenn.edu, yuetaili@uw.edu, xyue2@andrew.cmu.edu

How does finetuning on one perception task affect zero-shot performance on other perception tasks?

Experimental Setup



Experimental Setup



Experimental Setup

Pretrained VLM

LoRA Finetuning

Finetuning on one task

Counting 🔢

Evaluate

Repeat for all tasks

Evaluate on all other tasks

BLINK Benchmark

13 Perception Tasks

Art Style🎨, Counting🔢, Forensic Detection 🕵️, Jigsaw🧩,
Functional Correspondence🔗, Spatial Reasoning 📐,
Multi-view Reasoning👀, Relative Depth 📏, Visual Similarity🖼️,
Object Localization🎯, Visual Correspondence 🪢 ,
Relative Reflectance 💡 , Semantic Correspondence 🧠

Experimental Setup

Pretrained VLM

LoRA Finetuning

Finetuning on one task

Counting 🔢

Evaluate

Repeat for all tasks

Evaluate on all other tasks

BLINK Benchmark

13 Perception Tasks

Art Style🎨, Counting🔢, Forensic Detection 🕵️, Jigsaw🧩,
Functional Correspondence🔗, Spatial Reasoning 📐,
Multi-view Reasoning👀, Relative Depth 📏, Visual Similarity🖼️,
Object Localization🎯, Visual Correspondence 🪢 ,
Relative Reflectance 💡 , Semantic Correspondence 🧠

Models

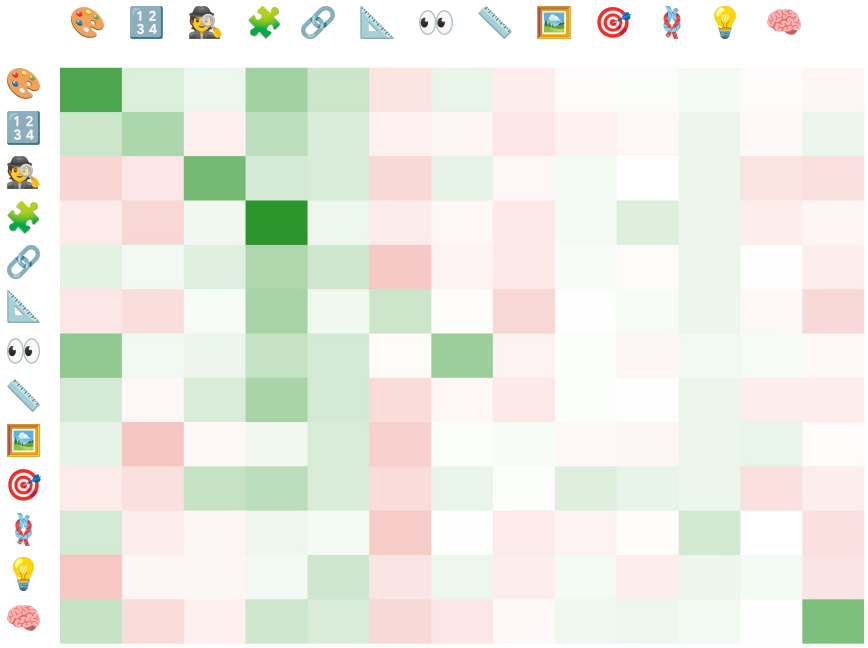
Qwen2.5-VL

3B, 7B and 32B
(across 4 seeds)

Task Transfer Matrix

Evaluate on target tasks

Finetune on source task



Positive transfer

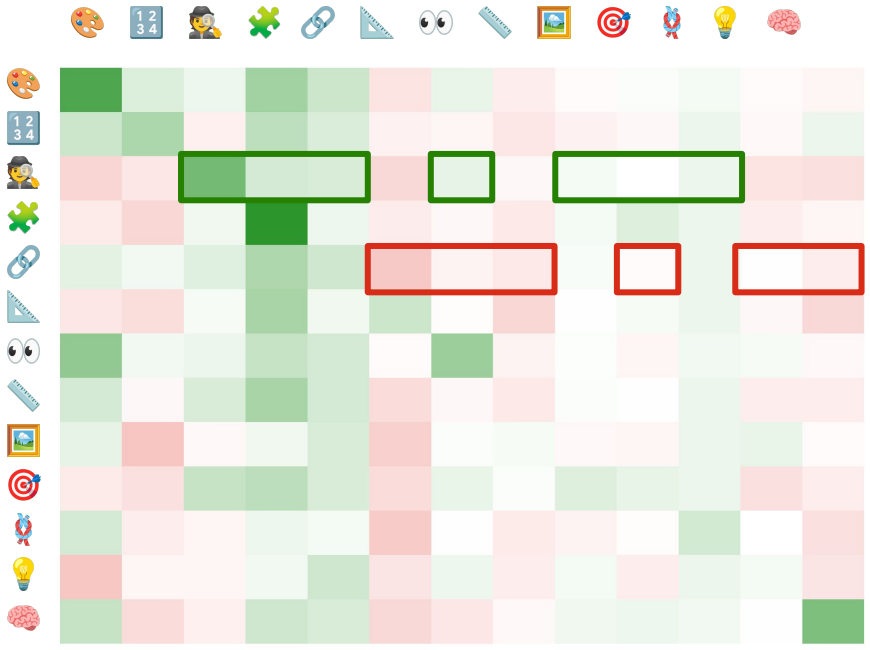
Transferability Score

Negative transfer

Task Transferability

Evaluate on target tasks

Finetune on source task

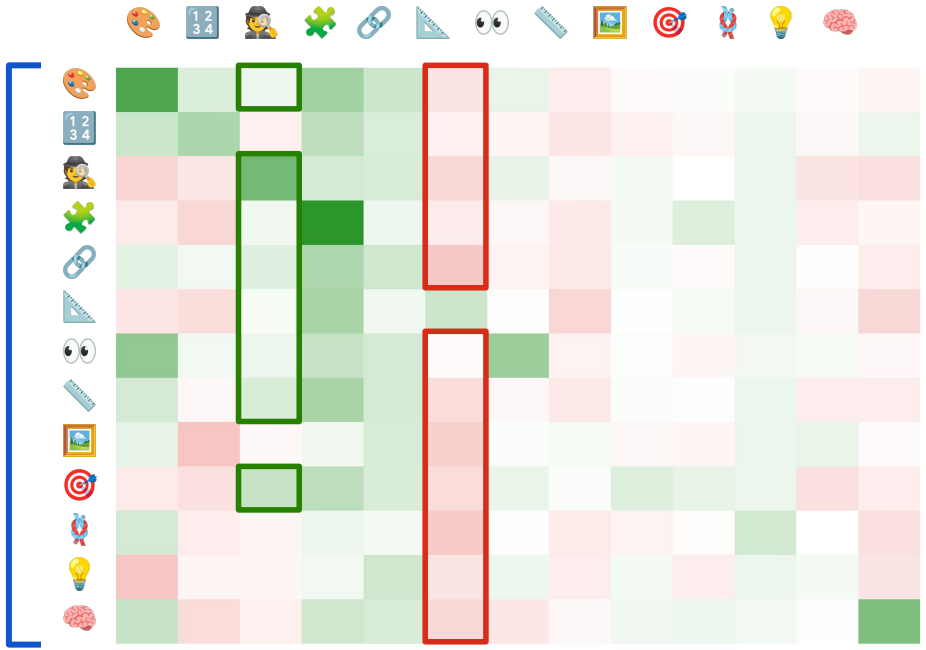


Aggregation of positive transferability scores and negative transferability scores for a source task is termed as its, respectively, positive and negative **Task Transferability**

Task Malleability

Evaluate on target tasks

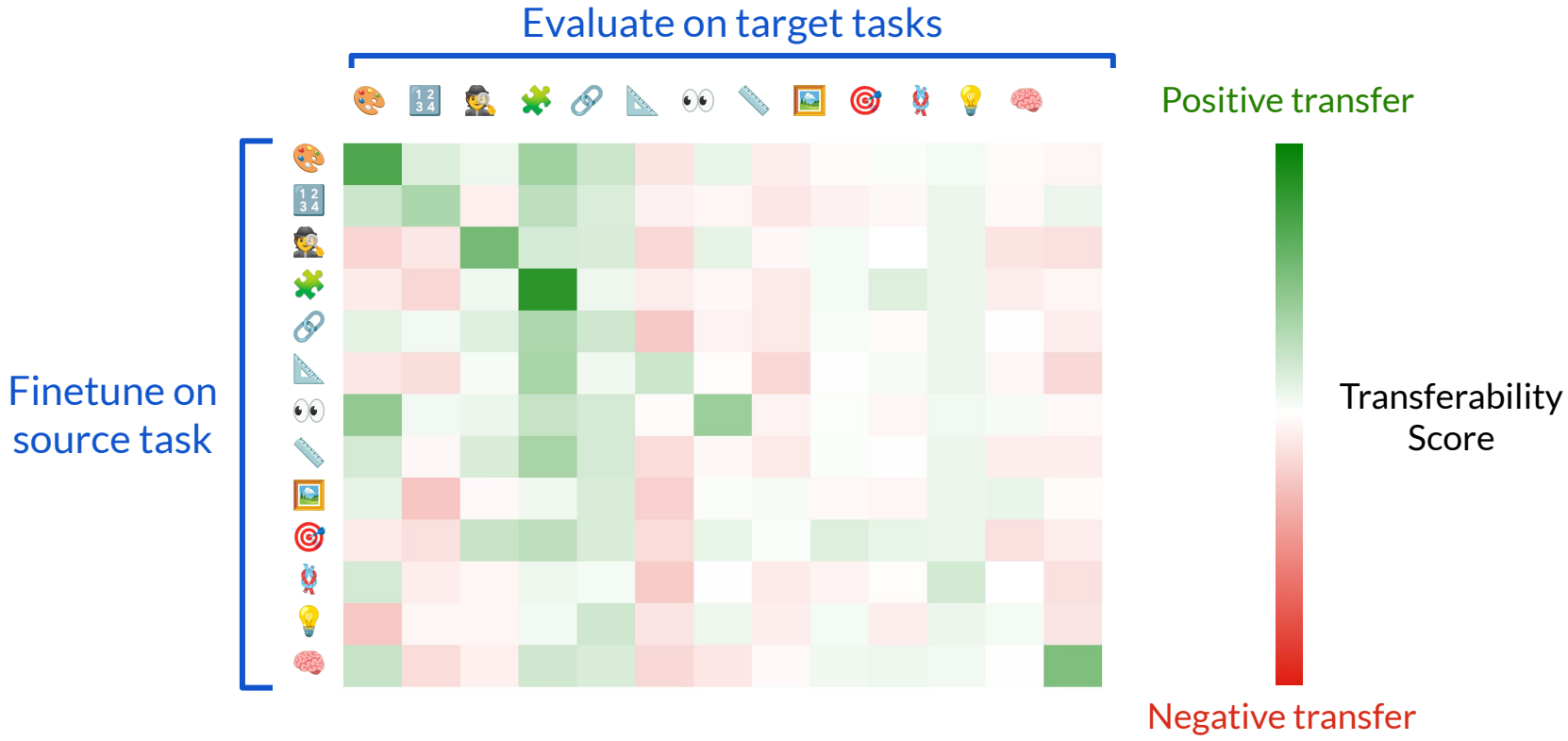
Finetune on source task



Helps measure how sensitive a target task is to finetuning on other tasks

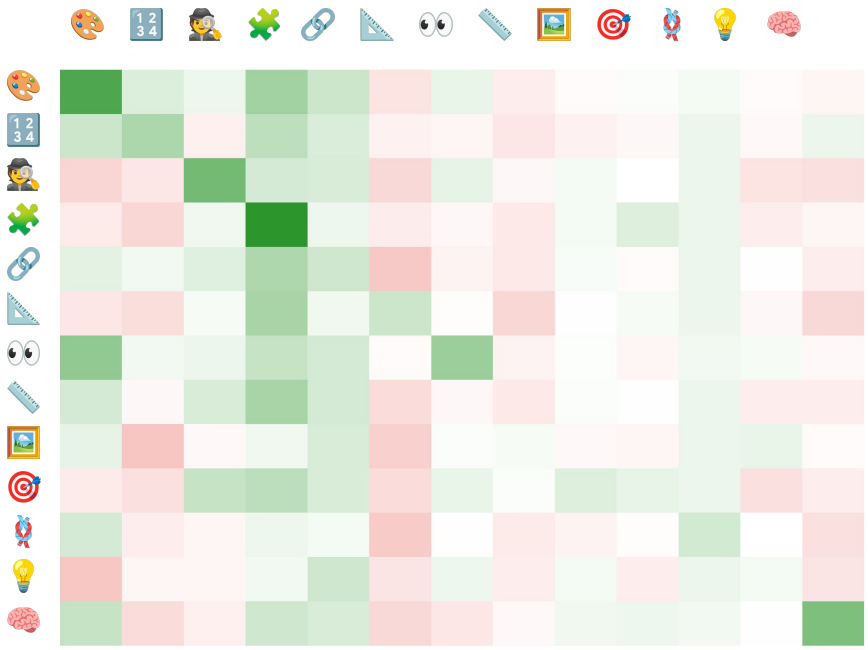
Aggregation of positive transferability scores and negative transferability scores for a target task is termed as its **Task Malleability**

Task Transfer Matrix



Task Transfer Matrix

What metric should be used for quantifying the transferability score?



Positive transfer

Transferability Score

Negative transfer

Why Raw Accuracy Gain Fails

Toy Example:

Task	Baseline Accuracy	After Finetuning	Raw Accuracy Gain
A	90%	93%	+3
B	40%	45%	+5

❓ Which transfer is actually stronger?
Raw gain can be misleading!

Perfection Gap Factor (PGF)

For a source task i and target task j:

$$\mu_{i \to j} = \frac{\text{Acc}(\mathcal{M}(T_i), T_j) - \text{Acc}(\mathcal{M}, T_j)}{U_j - \text{Acc}(\mathcal{M}, T_j)} = \frac{\text{Change in accuracy after finetuning}}{\text{Remaining headroom to upperbound accuracy}}$$

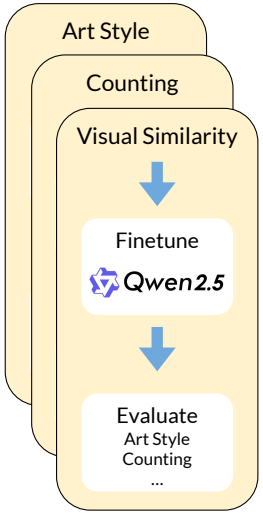
- $\text{Acc}(\mathcal{M}, T_j)$: zero-shot accuracy of model M on target task j
- $\text{Acc}(\mathcal{M}(T_i), T_j)$: accuracy after finetuning on source task i
- U_j : upperbound accuracy on task j

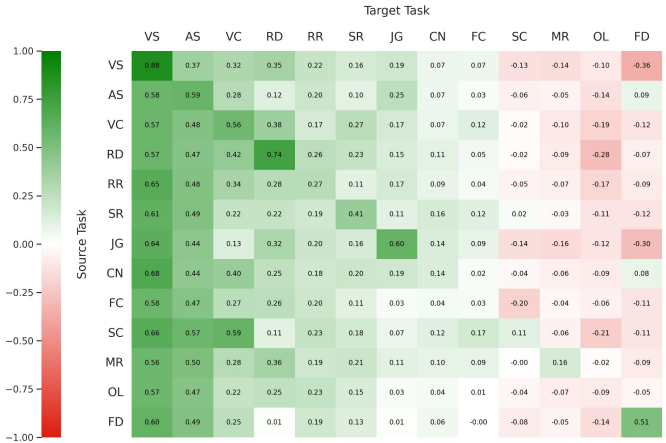
Interpretation

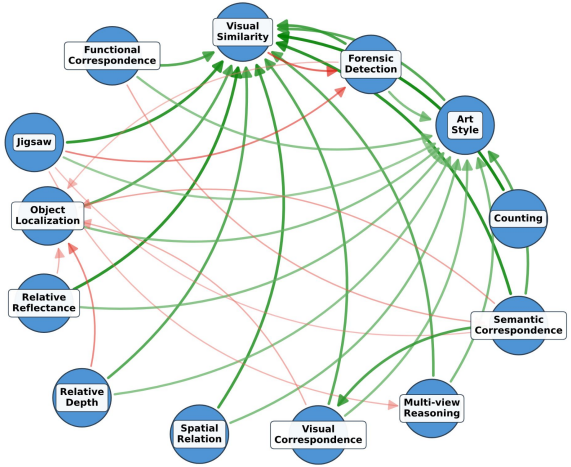
- **PGF > 0 : positive transfer**
- **PGF = 0 : no change**
- **PGF < 0 : negative transfer**
- **Bounded in [-infinity, 1]**

Task Graph

Using the defined setup and metrics, we obtain multiple task transfer matrices which we average out across seeds and create task graphs which have each node as a task and edges as their relation



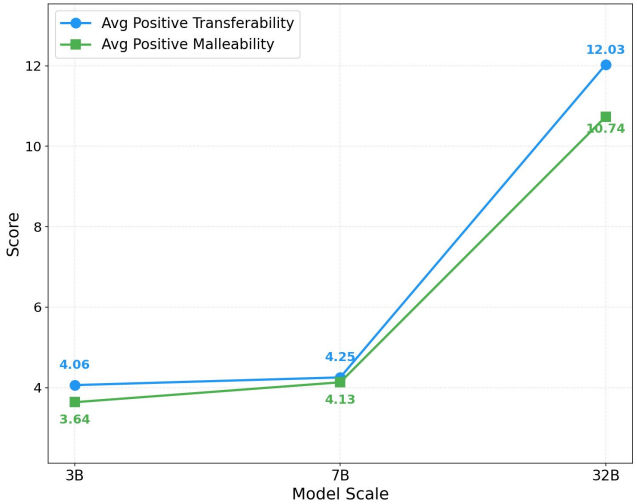




Key Takeaways

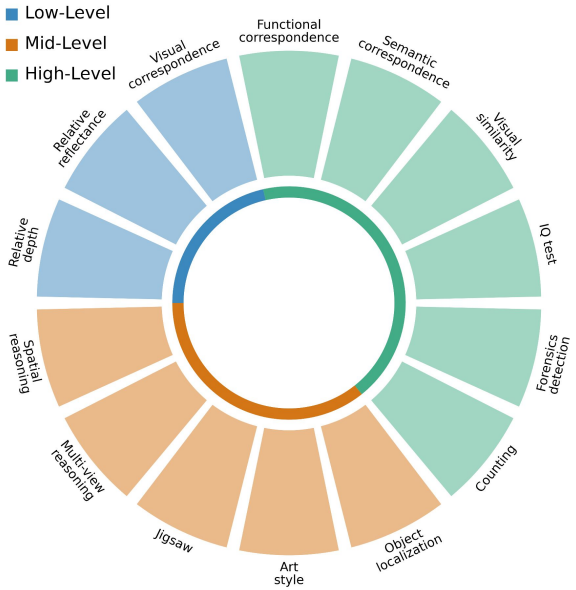
Model Scale vs Transfer

The magnitude of positive transferability and malleability increases with model size



Task Transfer across Categories

Zooming into from model scale, we also look into the trends based on task categorization

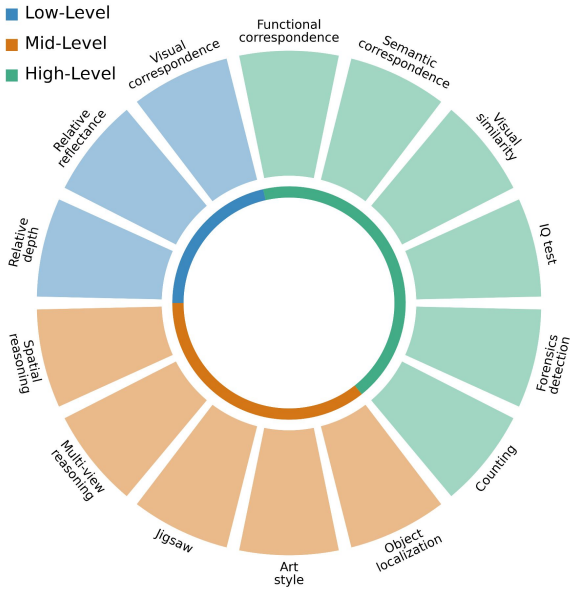


BLINK Benchmark provides a task categorization into

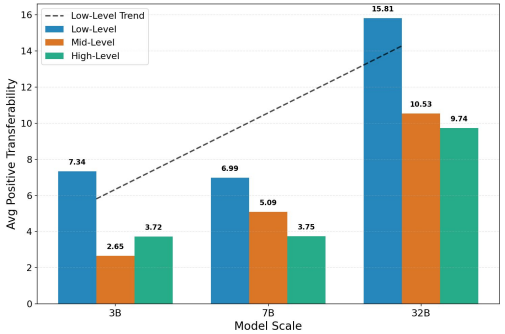
- Low-level
- Mid-level
- High-level

Task Transfer across Categories

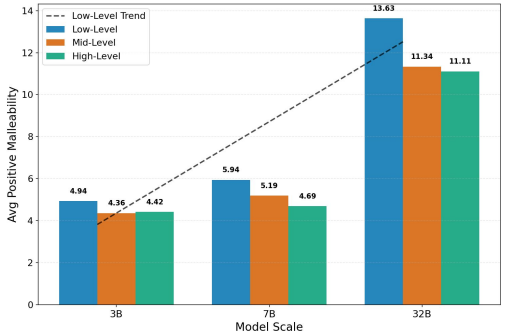
Low-level tasks are highly positively transferable and malleable. Finetuning on low-level tasks is beneficial compared to mid and high-level tasks



Higher positive transferability



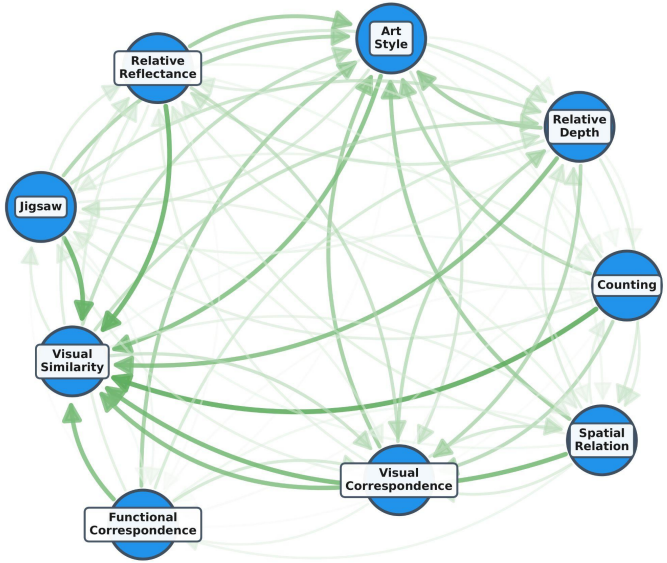
Higher positive malleability

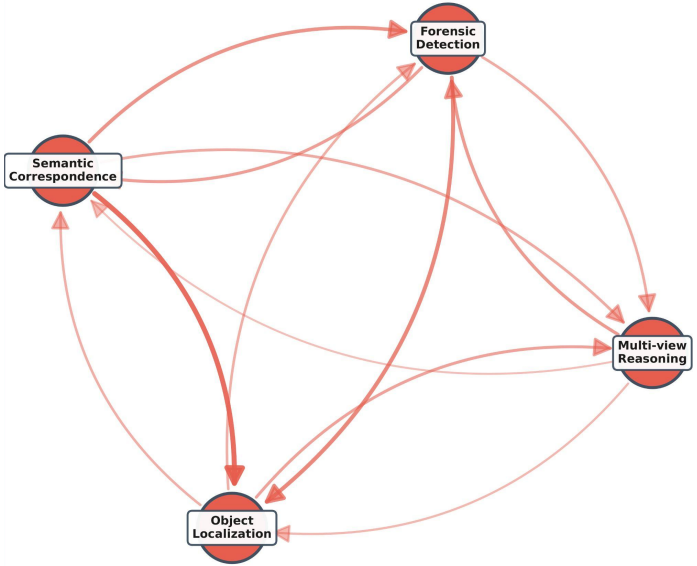


Increasing model size from 3B to 32B

Cliques of Cooperation

Tasks form cliques of mutually beneficial and mutually deteriorative clusters





Task Personas

Donors
Helps many other tasks

Pirates
Hurts many other tasks

Sponges
Easily improved by other tasks

Sieves
Easily degraded by other tasks

Task Personas

Donors
Helps many other tasks

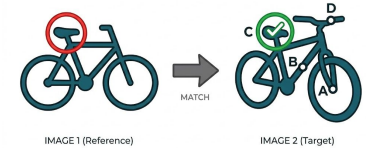
Pirates
Hurts many other tasks

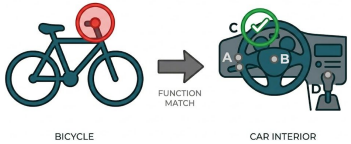
Sponges
Easily improved by other tasks

Sieves
Easily degraded by other tasks

Semantic Correspondence 🧠

Functional Correspondence 🔗





Task Personas

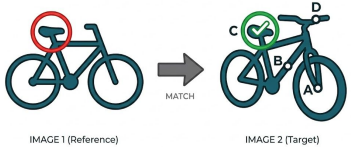
Donors
Helps many other tasks

Pirates
Hurts many other tasks

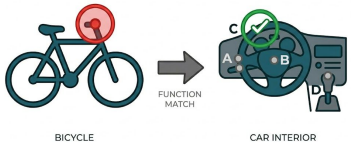
Sponges
Easily improved by other tasks

Sieves
Easily degraded by other tasks

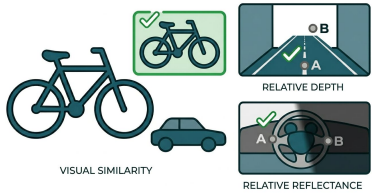
Semantic Correspondence 🧠



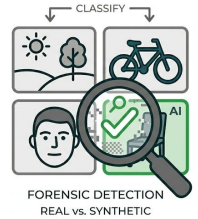
Functional Correspondence 🔗



Visual Similarity 🖼️
Relative Depth 📏
Relative Reflectance 💡



Forensic Detection 🕵️

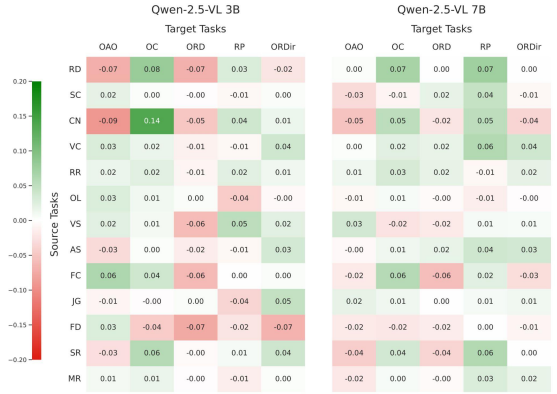


Transfer to Video Tasks

Image-level perception tasks induce positive transfer to video-based tasks as well

Finetune on Image Tasks in the BLINK Benchmark

Evaluate on Video Tasks in VSI-Bench



Similar task persona trends appear in the video domain as well!

Practical Implications

Dataset Curation

Better dataset selection, identifying foundational tasks, compute efficient training by avoiding harmful finetuning

Learning Paradigms

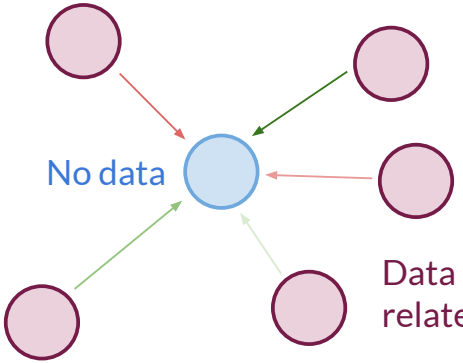
Safer and efficient continual learning, curriculum design and task ordering, cross-modal generalization

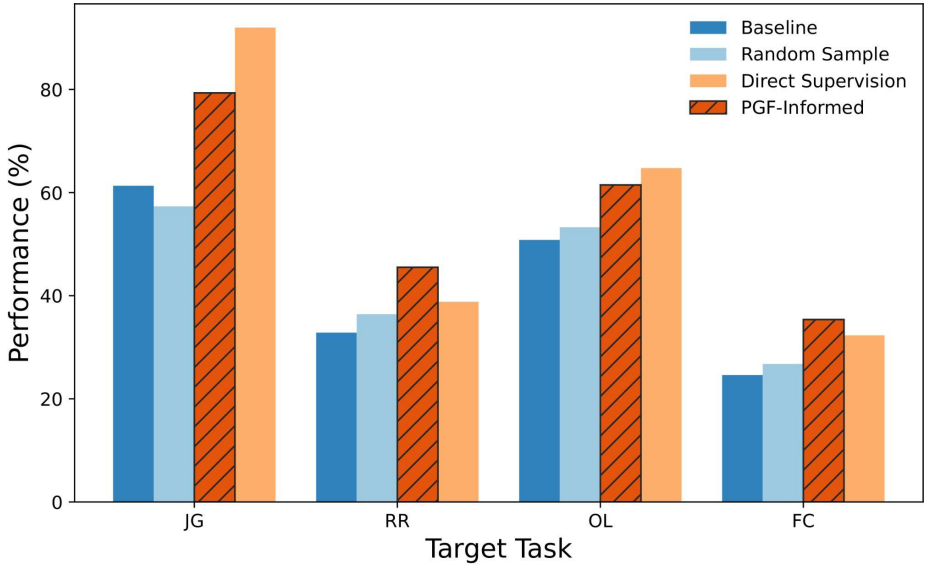
Benchmark Design and Evals

Detecting redundant tasks, evaluating robustness across capabilities, more principled synthetic data generation

PGF-Guided Dataset Selection

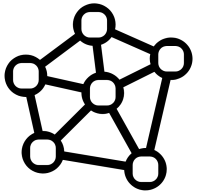
Consider the scenario where we want to optimize performance on some task for which no training data is available. Instead we have access to datasets from several related tasks.





When lacking supervised data, PGF-informed data selection can inform alternative dataset designs which can match and even exceed performance of direct finetuning

Key Contributions



Systematic study on broad suite of perception tasks

Uncover consistent structural properties of transfer, including scale-dependent trends, task categorization and task clusters

PGF enabled normalized cross-task analysis

Perfection Gap Factor (PGF) helps us conduct this analysis, by normalizing tasks across heterogeneous difficulty levels

Downstream Applications

Our analysis has several practical implications, ranging from better dataset selection to learning paradigms

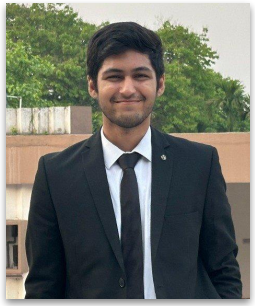
Project Page

Paper Link

Thank you!

Feel free to ask questions

Bhuvan Sachdeva



Karan Uppal

Abhinav Java

Vineeth N. B.

Appendix

Behaviour of PGF