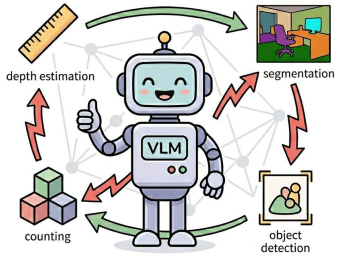
Microsoft Research



CVPR JUNE 3-7, 2026 DENVER COLORADO

Understanding Task Transfer in Vision-Language Models

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

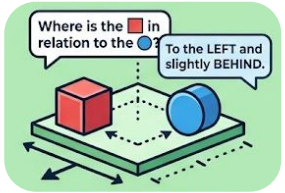
Microsoft Research India

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 🏔️

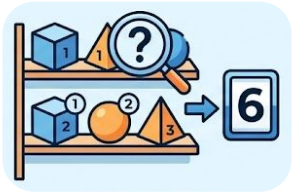
Relative Depth 📏

Object Localization 🎯

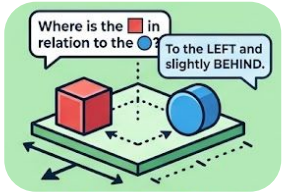
....

Finetuning VLMs on Perception Tasks

Perception tasks humans find natural are quite hard for current VLMs



Counting 🔢



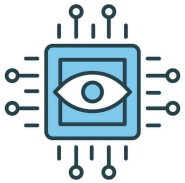
Spatial Reasoning 🏔️

Relative Depth 📏

Object Localization 🎯

....

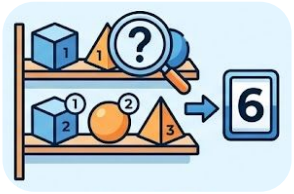
Current Scenario



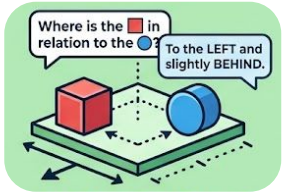
VLM → Finetune on one task 📐 → Good results on 📐⬆

Finetuning VLMs on Perception Tasks

Perception tasks humans find natural are quite hard for current VLMs



Counting 🔢



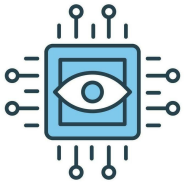
Spatial Reasoning 🏔️

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!

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

Experimental Setup

Pretrained VLM

LoRA Finetuning

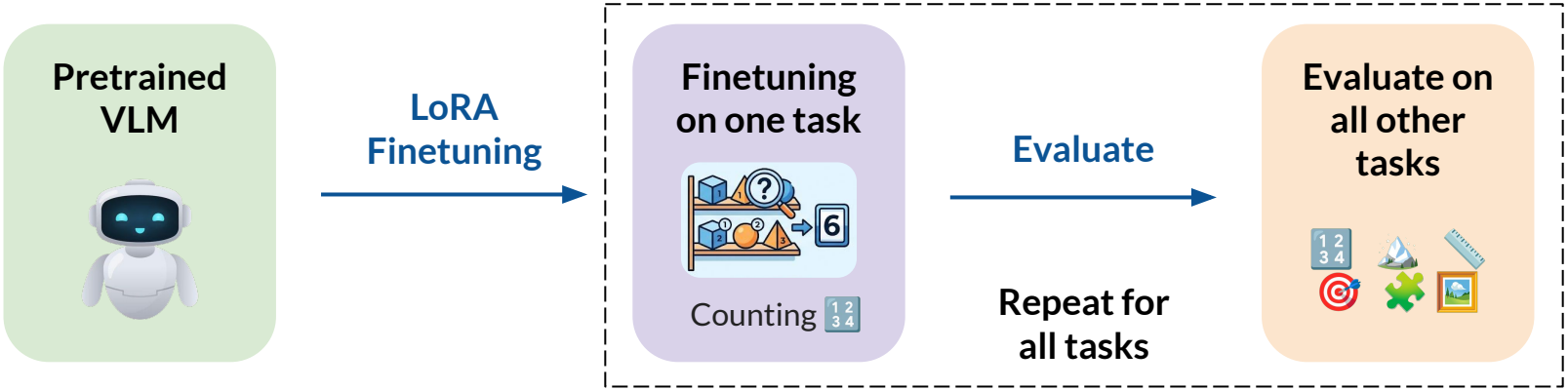
Finetuning on one task

Counting 🔢

Evaluate

Evaluate on all other tasks

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

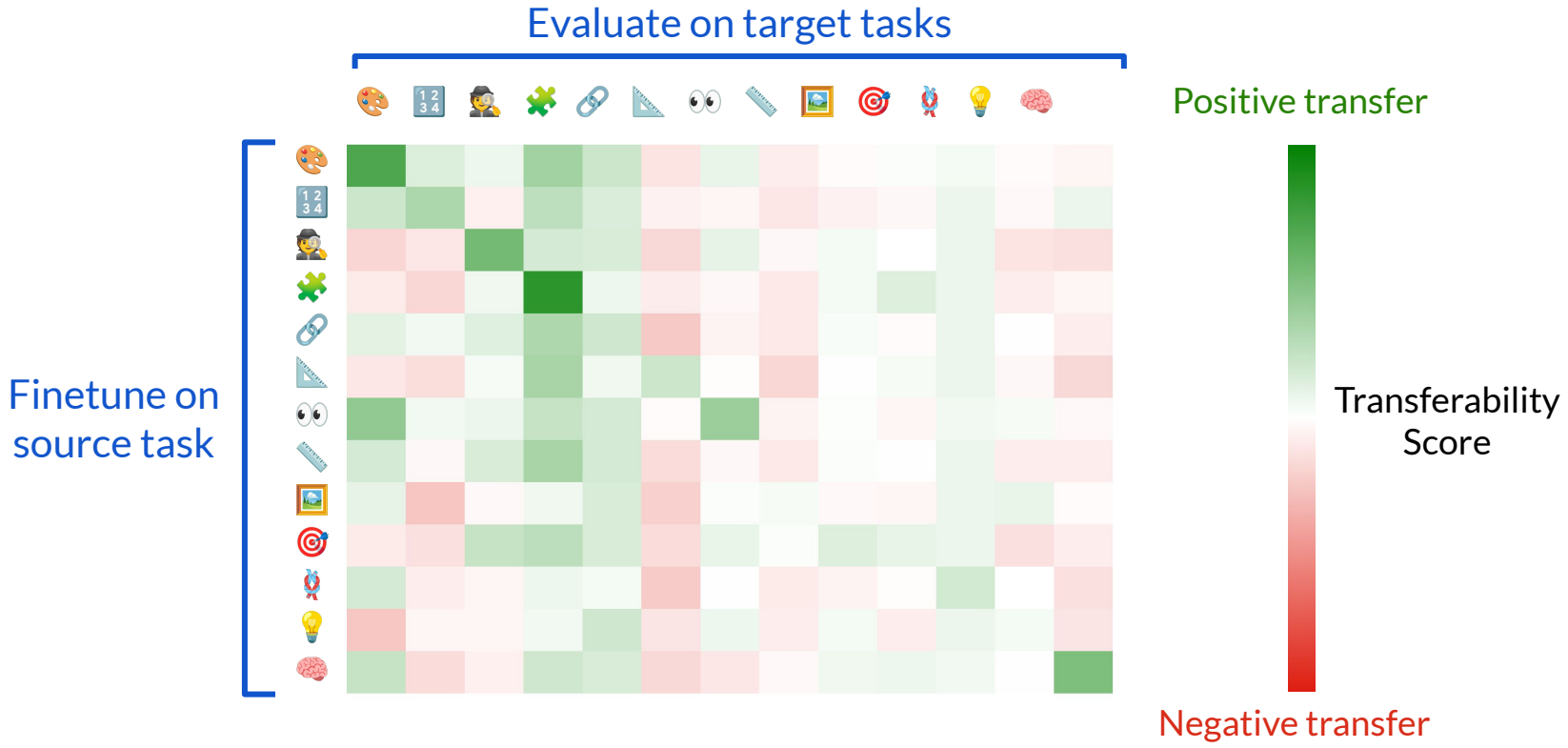
Art Style🎨, Counting🔢, Forensic Detection 🕵️, Jigsaw🧩,
Functional Correspondence🔗, Spatial Reasoning 📐,
Multi-view Reasoning👀, Relative Depth 📏, Visual Similarity🖼️,
Object Localization🎯, Visual Correspondence 🪢 ,
Relative Reflectance 💡 , Semantic Correspondence 🧠

Models

Qwen2.5-VL

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

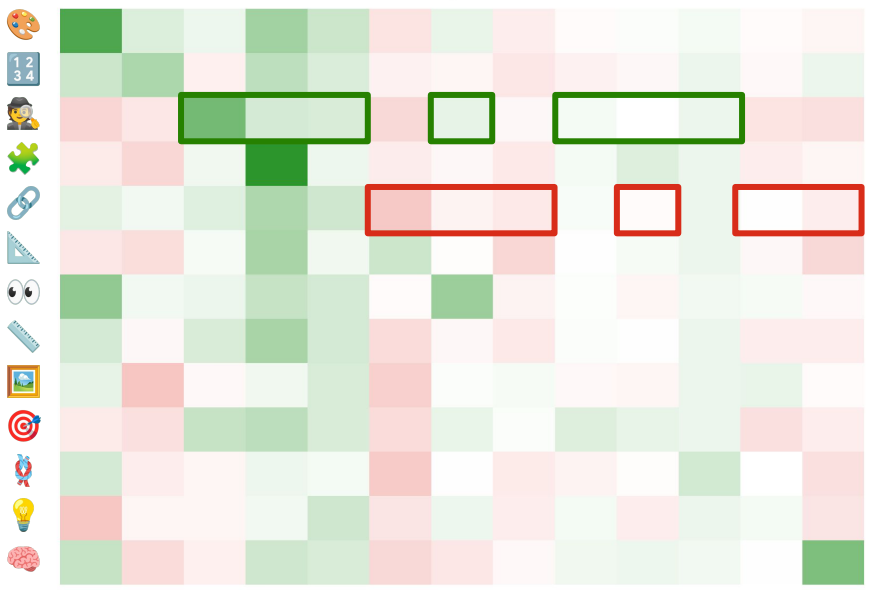
Task Transfer Matrix



Task Transferability

Evaluate on target tasks

Finetune on source task

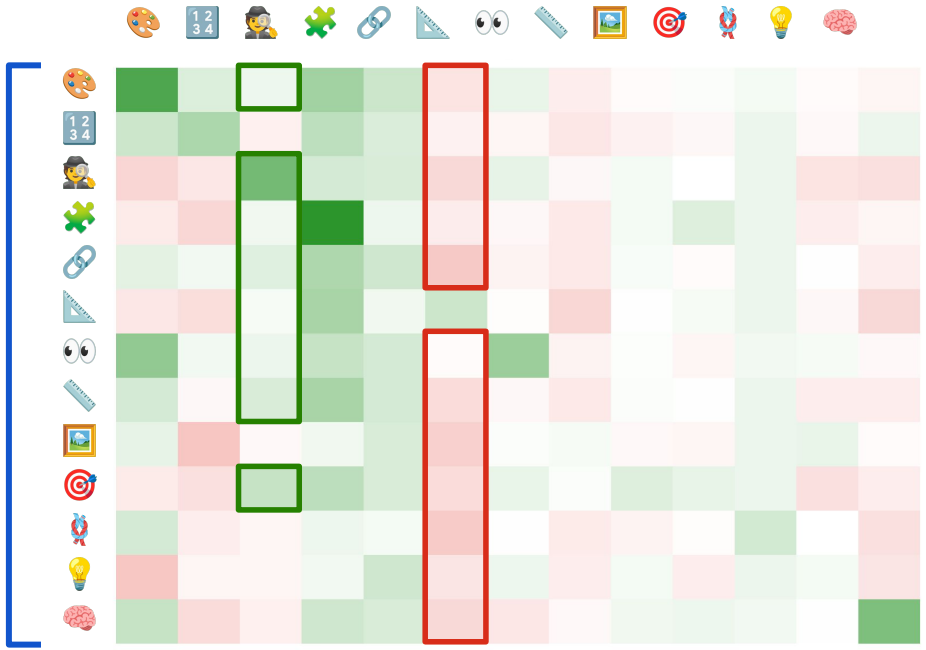


Aggregation of positive transferability scores and negative transferability scores for a source task is termed as its, respectively, positive and negative **Task Transferability**

Task Malleability

Evaluate on target tasks

Finetune on source task



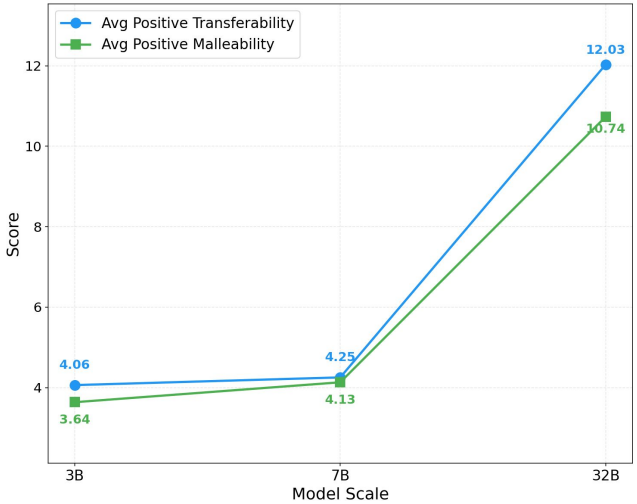
Helps measure how sensitive a target task is to finetuning on other tasks

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

Key Takeaways

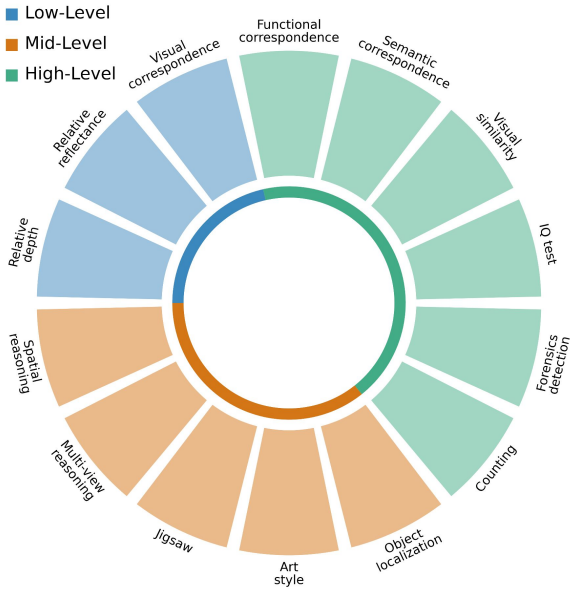
Model Scale vs Transfer

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

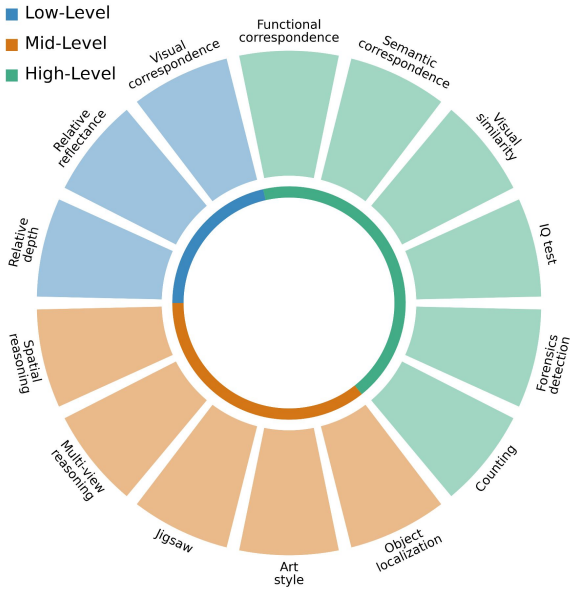


BLINK Benchmark provides a task categorization into

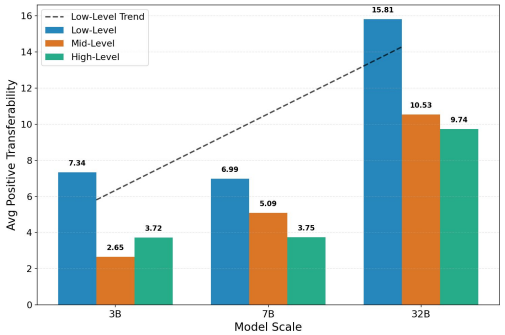
- Low-level
- Mid-level
- High-level

Task Transfer across Categories

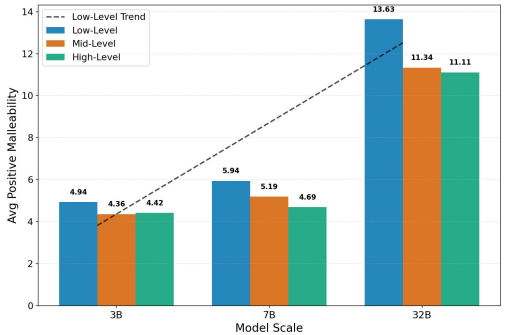
Low-level tasks are highly positively transferable and malleable. Finetuning on low-level tasks is beneficial compared to mid and high-level tasks



Higher positive transferability



Higher positive malleability



Increasing model size from 3B to 32B

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

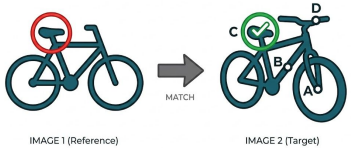
Easily improved by other tasks

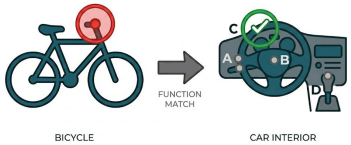
Sieves

Easily degraded by other tasks

Semantic Correspondence 🧠

Functional Correspondence 🔗



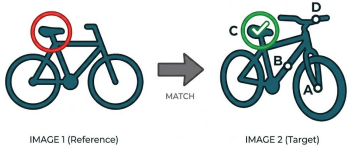


Task Personas

Donors

Helps many other tasks

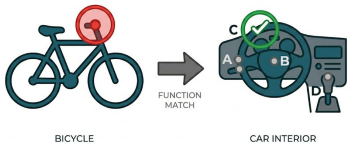
Semantic Correspondence 🧠



Pirates

Hurts many other tasks

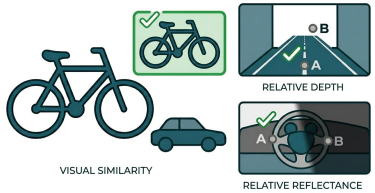
Functional Correspondence🔗



Sponges

Easily improved by other tasks

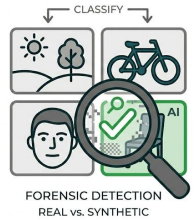
Visual Similarity🖼️
Relative Depth📏
Relative Reflectance 💡



Sieves

Easily degraded by other tasks

Forensic Detection🕵️



Practical Implications

Dataset Curation

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

Learning Paradigms

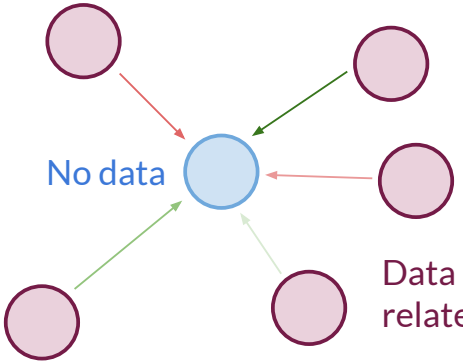
Safer and efficient continual learning, curriculum design and task ordering, cross-modal generalization

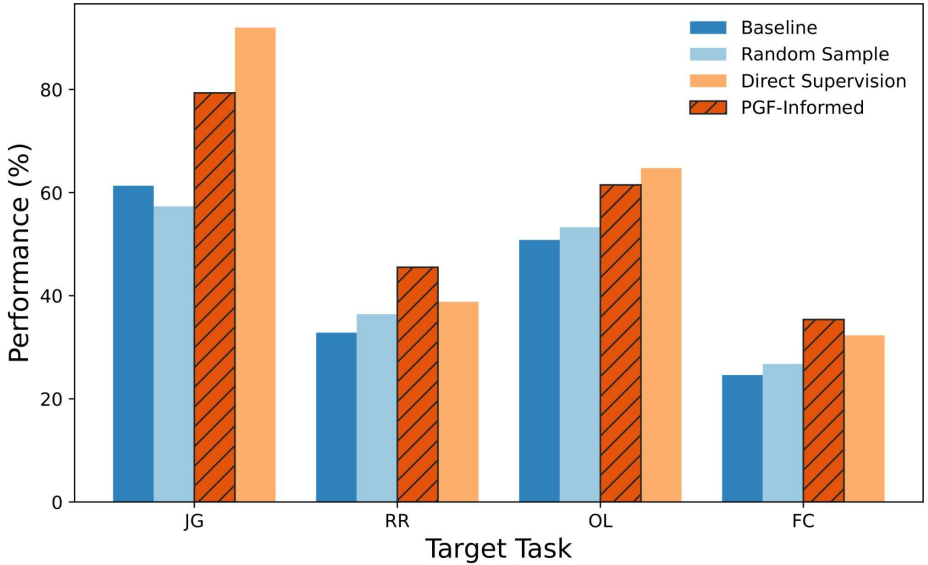
Benchmark Design and Evals

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

Guided Dataset Selection

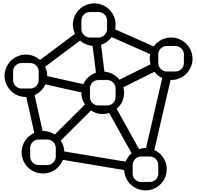
Consider the scenario where we want to optimize performance on some task for which no training data is available. Instead we have access to datasets from several related tasks.





When lacking supervised data, transfer-informed data selection can give alternative dataset designs which can match & 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

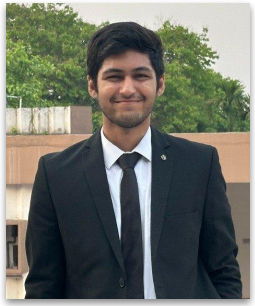
Project Page

Paper Link

Thank you!

Feel free to ask questions

Bhuvan Sachdeva



Karan Uppal

Abhinav Java

Vineeth N. B.