Perle + DataSeeds.AI White Paper: Peer-Ranked Precision

How a Human-Verified Image Dataset Outperforms Traditional Vision Benchmarks

Sajjad Abdoli, PHD - Founding AI Scientist, Perle
Freenam Lewin - Founder, Brickroad
Gediminas Vasiliauskas - AI/ML Engineer, Zedge

Perle DataSeeds.AI Brickroad

Index

03 Introduction

05 Human-in-the-Loop Annotation at Scale: DSD Segmentation & Annotation

07 Semantic Segmentation

09 Dataset Analysis

12 Benchmarking the DSD

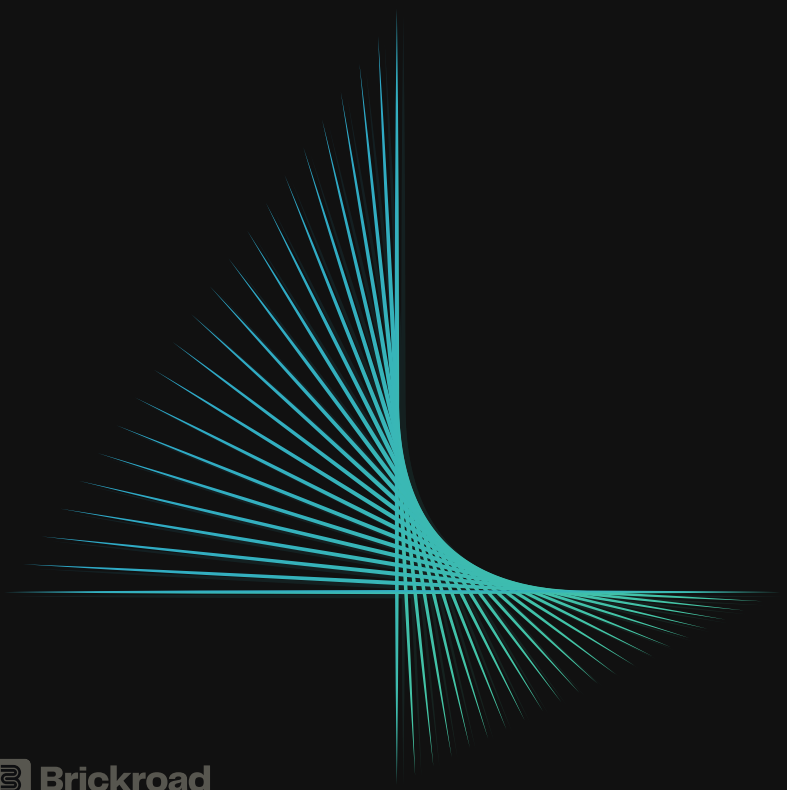
14 Baseline Evaluation: AWS Rekognition vs. DSD

16 Fine-Tuned Model Evaluation: LLAVA-NEXT and BLIP2

20 Conclusion

23 Elevate Your AI

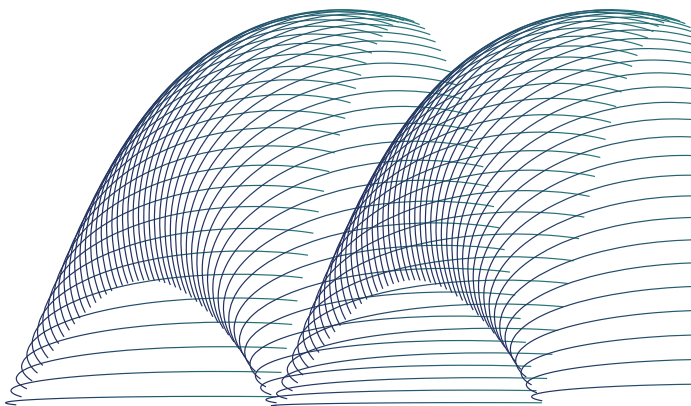
25 References



Introduction: A Paradigm Shift in AI

A silent revolution is underway in AI. As performance gains plateau under traditional model-centric approaches to AI optimization, attention has turned towards an oft-neglected variable: the data itself. This “data-centric” approach has gained prominence through the work of researchers like Bhatt et al. (2024) and others who have gone great lengths to demonstrate the value of dataset quality, structure, and richness on model output and resource savings.

Yet, despite this evidence, widespread adoption of data-centric AI model training remains hamstrung by economic and political roadblocks. Most of the AI ecosystem still relies on legacy datasets like ImageNet and OpenImages, which suffer from noisy crowd-sourced labels, low annotation fidelity, and limited task-specific utility. These datasets, while historically valuable, are poorly suited for emerging use cases like multimodal reasoning, fine-grained visual analysis, and generative AI.



To address this gap, Perle, DataSeeds.Ai (a Zedge company), and partnered Brickroad to produce a new foundational dataset: the **DataSeeds.AI Sample Dataset (DSD)**.

Derived from the GuruShots photography game, which is part of Zedge, the DSD comprises 7,772 expert-annotated and peer-ranked images drawn from a catalog of over 100 million licensable photos.

The DSD introduces structured human judgement, dense annotation, and high-aesthetic quality into one coherent framework, paving the way for precision AI. Multiple tiers, bridging the gap between perception and machine understanding.

In the process of developing the DSD, to validate and benchmark its effectiveness, the authors choose to publish, along with the dataset, their findings in the form of a comprehensive research paper, as well as the weights and code.

Research Paper

DSD Dataset

LLaVA weights

BLIP weights

Code scripts

Human-in-the-Loop Annotation at Scale: DSD Segmentation & Annotation

High-quality annotation should be approached as a necessity for AI labs looking to advance the state of computer vision modeling.

Inspired by recent advances in semi-supervised pipelines such as NVIDIA's Describe Anything Model (DAM) (Lian et. al. 2025), Perle developed a multi-tier human-in-the-loop annotation pipeline for the DSD.

This approach integrates structured human input with scalable machine assistance, delivering a dataset that balances annotation quality and efficiency.

Each DSD image is annotated with:

- Pixel-level segmentation masks
- Three tiers of human-generated text:

1. A concise title
2. A 15+ word image description
3. A 20–30 word technical scene analysis

This structure delivers compositional insight, photographic context, and aesthetic interpretation—providing critical data for training models on human-like reasoning tasks.

Image ID: c680f5c5-9543-4953-af96-a7c123609673

Image Information

Title: Living Room in a house

Description: Living room with couch in front of the bay windows and coffee table in front of it and under the coffee table there is a rug. And beside the couch, from the two sides, there are two tables with lamps on them and

Scene: The photo was captured from the point of view angel for a living room with mixed classic and modern furniture in it with a multicolor color palette, and it

Segmented Objects (26 masks)

Living Room	Rug	Couch
Coffee Table	Table	Coffee Table
Mirror	Painting	Candle holder
Clock	Side table	Table lamp
lighting fixture	Baggage	Baggage
Baggage	curtain	Bay Window
Bay Window	Bay Window	Cushions
Cushions	Bottles	flower vase
flower vase	Lamps	

All Labels (30 total)

• Home Decor	• Architecture	• Building	• Furniture
• Indoors	• Living Room ★	• Room	• Table ★
• Couch ★	• Coffee Table ★	• Rug ★	• Interior Design
• Window	• Candle	• Art	• Painting ★
• Bay Window ★	• Cushion	• curtain ★	• Mirror ★
• Candle holder ★	• Clock ★	• Side table ★	• Table lamp ★
• lighting fixture ★	• Baggage ★	• Cushions ★	• Bottles ★
• flower vase ★	• Lamps ★		

Semantic Segmentation

Unlike many datasets that rely on bounding boxes, the DSD adopts full semantic segmentation, providing pixel-level delineation of objects. Bounding boxes often include irrelevant background and lack the granularity needed for high-resolution scene understanding.

Our segmentation captures occlusion, overlap, and object contours with high precision, enabling improved spatial reasoning and feature extraction. This granularity is essential for applications such as generative modeling, AR/VR scene composition, and image editing, where spatial relationships and morphological accuracy are critical.

For example:

Title: Honeybee in the Field

Description: A flying honeybee approaches a cluster of red and yellow blooms, its delicate wings beating rapidly. The bee's tiny legs dangle, ready to land, while its fuzzy golden body glows in sunlight.

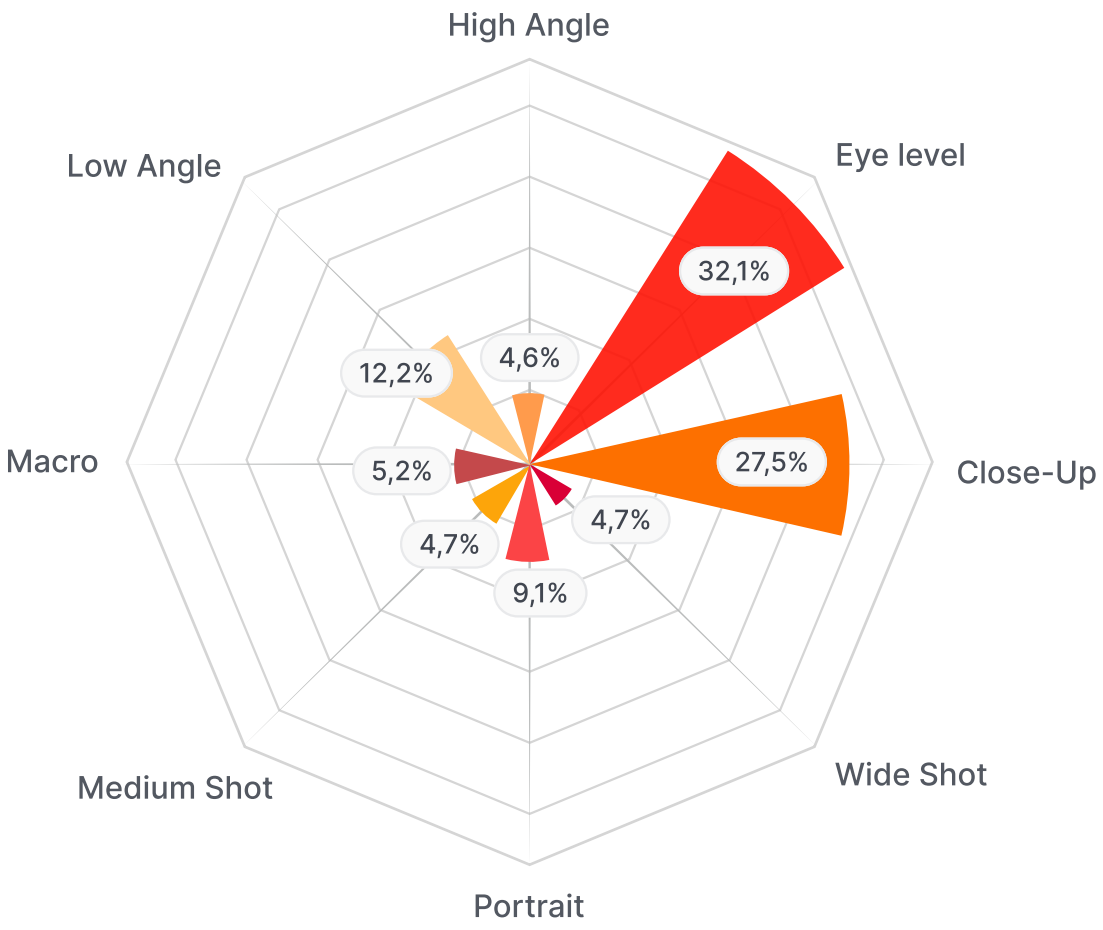
Scene Analysis: Close-up macro shot from a straight-on angle, providing an intimate view of the bee's details and the flower's textures. These annotations provide critical semantic and compositional details inaccessible to automated systems and are pivotal for enabling technical scene understanding in modern multimodal AI models.

Dataset Analysis

Testing on the DSD occurred across 10,610[1] meticulously annotated images. Visual analysis reveals a predominance of close-up and eye-level compositions, followed by portraits and wide-angle shots, indicating a preference for object-centric framing.

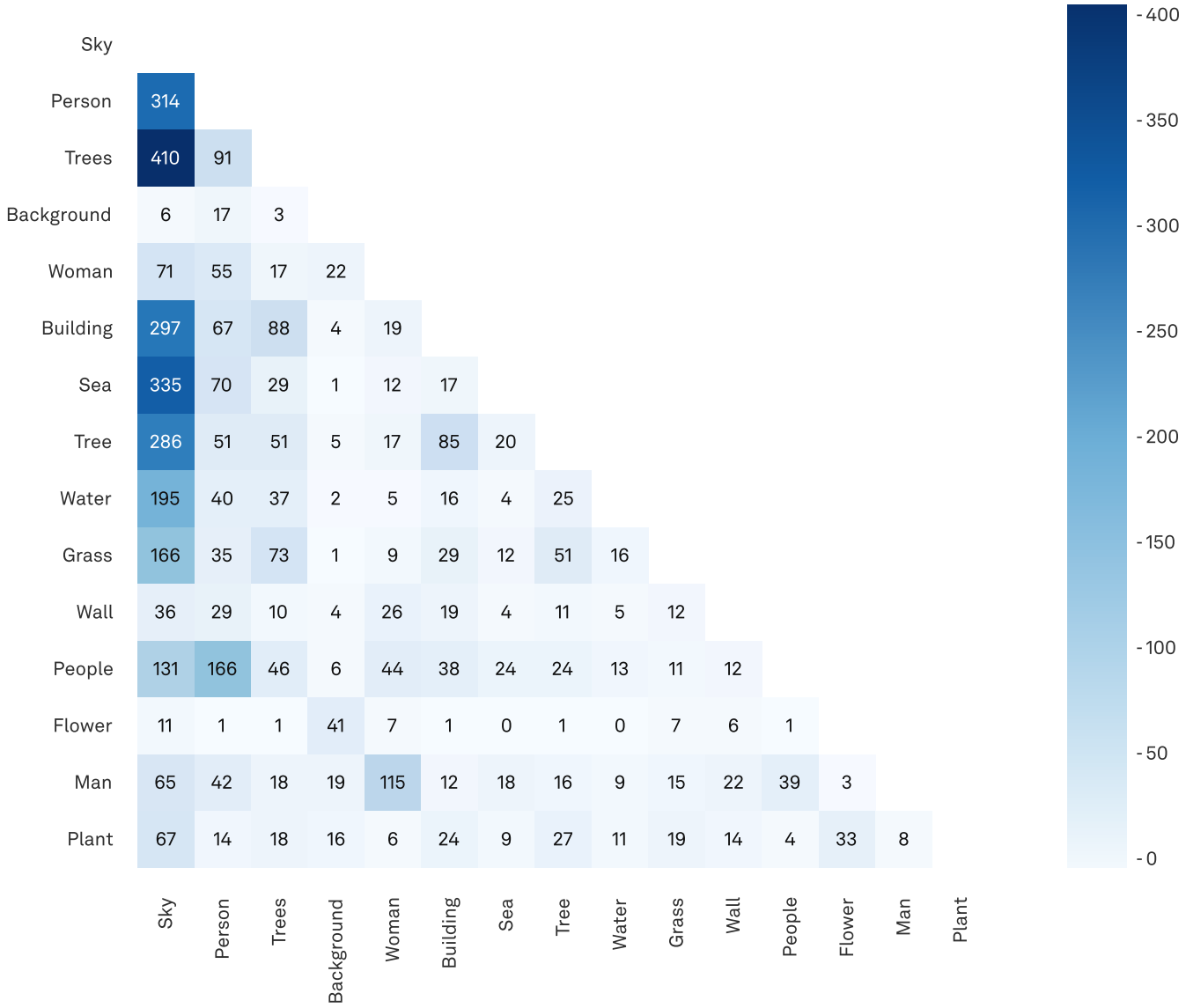
Frequent natural elements-sky, trees, grass-highlight a strong outdoor and environmental presence. Human-centric tags like person, man, and woman are also prevalent, enriching interpersonal and contextual cues.

[1] Experiments undertaken for this research were done on the DSD which contained 10,610 images. Prior to publication, in order to account for a wide range of multi-jurisdictional compliance considerations, we removed 2,767 images that contained subjects of a sensitive nature, including people's faces. The remaining 7,772 will be available for public usage at: https://huggingface.co/datasets/Dataseeds/DataSeeds.AI-Sample-Dataset-DSD



Semantic co-occurrence analysis shows, for example, frequent pairing of sky with trees and grass, and over 100 instances where man and woman appear together. These correlations support ontology development and semantic hierarchies, which can enhance model training by encoding label interdependencies.

Co-occurrence heatmap of the top 15 labels



Benchmarking the DSD

Objectives & Hypotheses

We established a rigorous framework to evaluate human-in-the-loop annotation value. Core hypotheses:

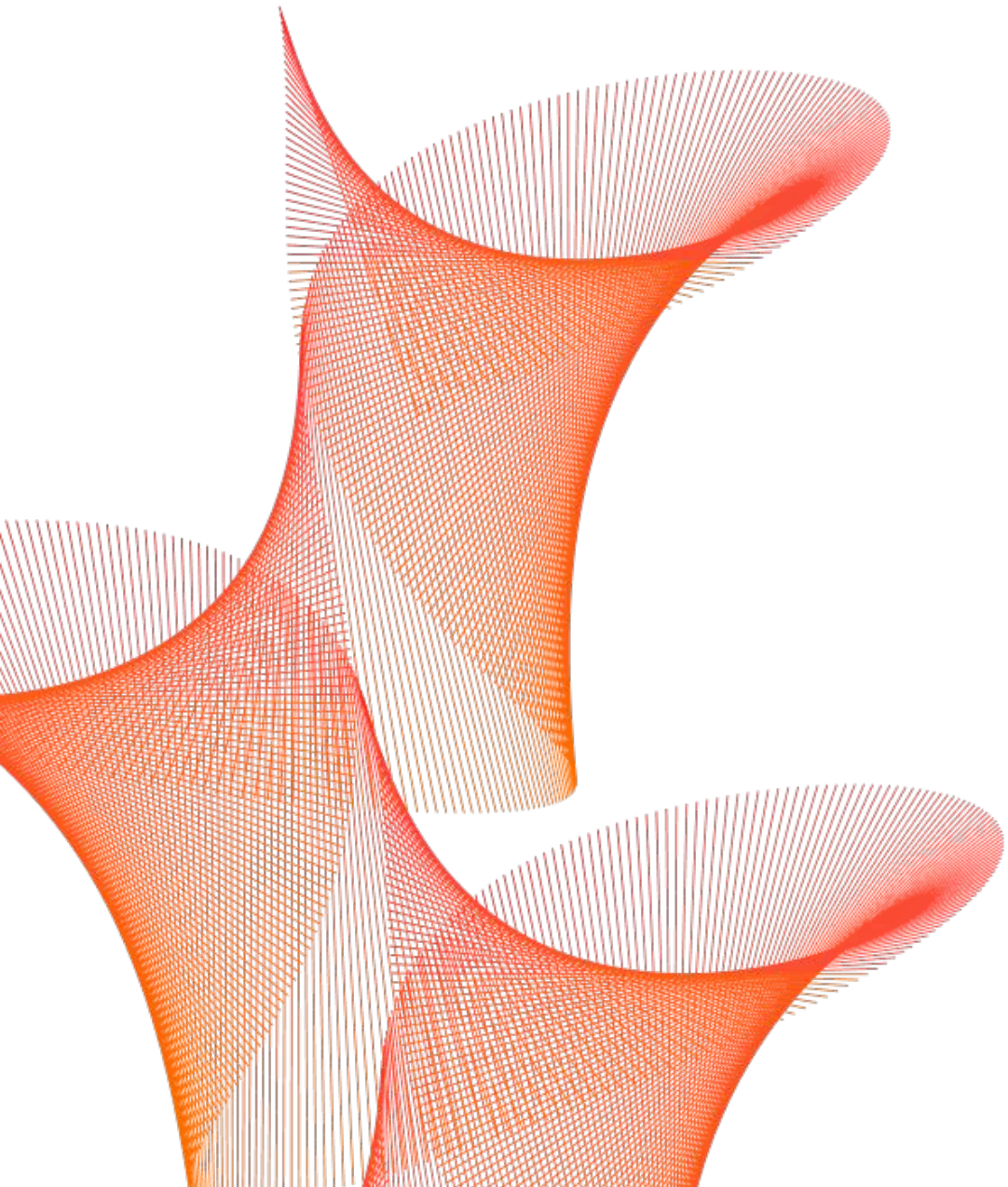
Core hypotheses:

> Commercial automated annotation systems exhibit semantic gaps, especially for technical and aesthetic attributes.

> Fine-tuning multimodal models with DSD improves technical scene analysis performance.

Central to this experiment was the hypothesis that exceptional image quality, combined with precise human-in-the-loop annotations, can materially enhance a model's ability to accurately contextualize and describe compositional and technical attributes, including depth of field, lighting techniques, and camera perspectives.

The findings herein prove that the exceptional visual fidelity of the DSD is essential for models seeking to become capable of professional-level scene analysis.



Two primary evaluations were conducted:

01 Baseline Evaluation: AWS Rekognition vs. DSD

Benchmarking Amazon Rekognition's label detection against our human-curated annotations revealed significant precision and recall gaps.

Rekognition underperformed notably in compositional and stylistic categories, underscoring the limitations of general-purpose systems for nuanced visual tasks.

Overall Performance Metrics of AWS Rekognition Label Detection

Metric	Value
Total Images Analyzed	10610
Average Precision	0.1359
Average Recall	0.4292
Average F1 Score	0.1913
Unique AI Labels	2335
Unique Human Labels	7926
Label Overlap	1503 (18.96% of human labels)

02 Fine-Tuned Model Evaluation: LLAVA-NEXT and BLIP2

LLaVA-NEXT

- Modular architecture with SigLIP vision encoder and Qwen2-0.5B language model.
- Fine-tuned via LoRA (rank=32, alpha=32), BF16 mixed-precision, 3 epochs on a single NVIDIA A100 40GB GPU.
- ~917MB model size; 56% parameters trained.
- Used scene description prompts for supervision with frequent validation and checkpointing.
- Achieved substantial improvements in detailed, stylistically accurate captions versus zero-shot baseline.

Performance Comparison: Base vs. Fine-tuned LLAVA-NEXT

Model	BLEU-4	ROUGE-L	BERTScore	CLIPScore
LLAVA-NEXT (Base)	0.0199	0.2089	0.2751	0.3247
LLAVA-NEXT (Fine-tuned)	0.0246	0.214	0.2789	0.326
Absolute Improvement	0.0048	0.0051	0.0039	0.0013
Relative Improvement (%)	24.09	2.44	1.4	0.41

BLIP2 (OPT 2.7B variant)

- Uses Q-Former bottleneck connecting vision encoder and language model.
- Fine-tuned with DSD's tiered annotations and technical prompts.
- Notable improvements in compositional and stylistic accuracy.

Performance Comparison: Base vs. Fine-tuned BLIP2 (OPT 2.7B)

Model	BLEU-4	ROUGE-L	BERTScore	CLIPScore
BLIP2 (Base)	0.001	0.126	0.0545	0.2854
BLIP2 (Fine-tuned)	0.047	0.242	-0.0537	0.2583
Absolute Improvement	0.046	0.116	-0.1082	-0.0271
Relative Improvement (%)	4600	92.06	-198.53	-9.49

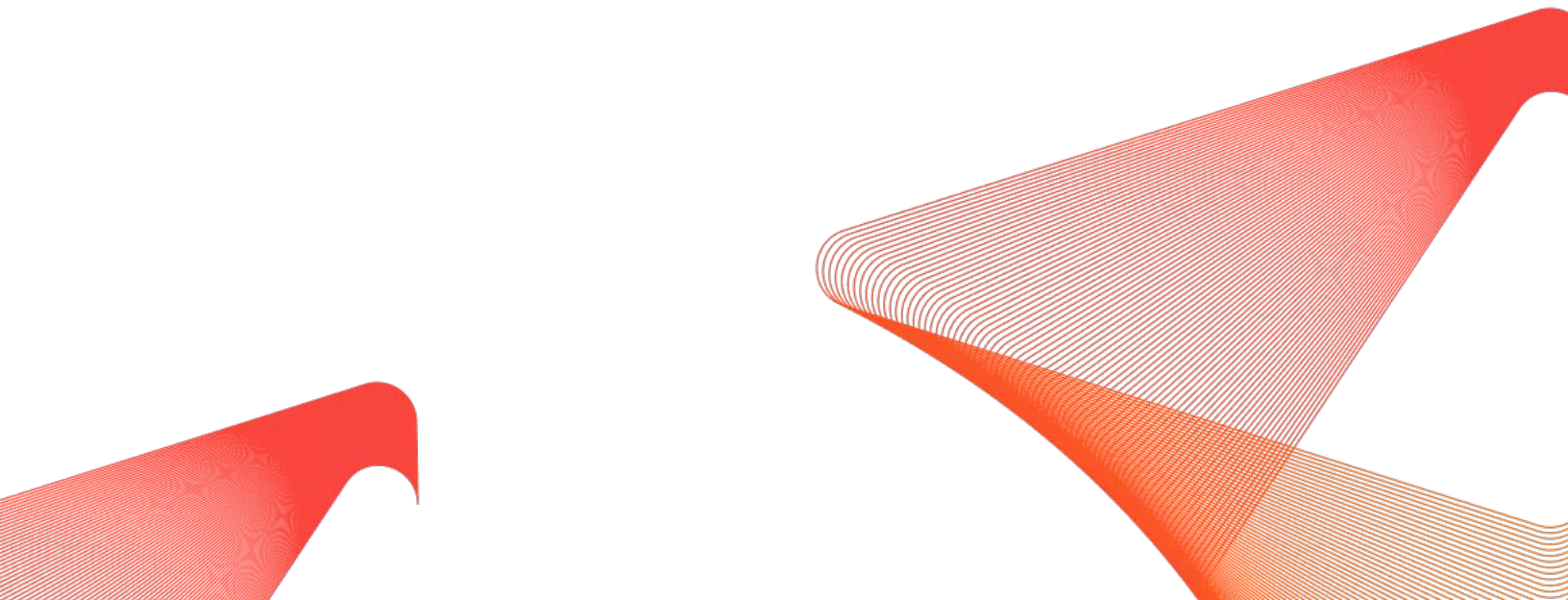
Both models saw substantial improvements in scene understanding, composition labeling, and stylistic fidelity.

Key Finding: LLaVA's modular alignment favored spatial accuracy; BLIP2's bottleneck architecture improved lexical expressiveness. Both benefited from the DSD's annotation density.

By comparing LLaVA and BLIP2, we observed how architectural differences mediate responsiveness to data quality:

- LLaVA-NEXT retained spatial fidelity, useful for camera angle and lighting classification.
- BLIP2 captured higher-level abstraction, generating stylistically rich captions.

Together, these findings validate the DSD's broad applicability across model paradigms.



Comparison of base and fine-tuned LLAVA-NEXT and BLIP2 model outputs

Name	Model
Reference	Human

Description

It's a close-up shot of a sleek pair of Sony WH-1000XM3 headphones resting on a dark surface. The color palette features matte black, deep gray, and subtle copper accents. The lighting is soft and directional, creating gentle highlights and shadows that emphasize the headphones' contours and premium texture. The mood is sleek, modern, and sophisticated, evoking a sense of quiet focus and premium quality.

Name	Model
LLAVA-NEXT	Base Model

Description

The image showcases a pair of sleek, black Sony headphones with orange accents. The headphones are resting on their sides against a dark background, highlighting the brand name "Sony" and model number "WH-1003". A close-up shot captures the ear cups in detail, showcasing ergonomic design and padding for comfort during extended use.

Name	Model
LLAVA-NEXT	Fine-tuned Model

Description

The image shows a pair of black Sony headphones with orange accents, resting on a dark surface. The camera angle is low and slightly above the headphones to provide an eye-level perspective. The lighting appears soft and diffused, likely from ambient light sources in a room or studio setting. There are no visible texts other than "WH-100XM3" printed near the ear cup area.

Name	Model
BLIP2	Base Model

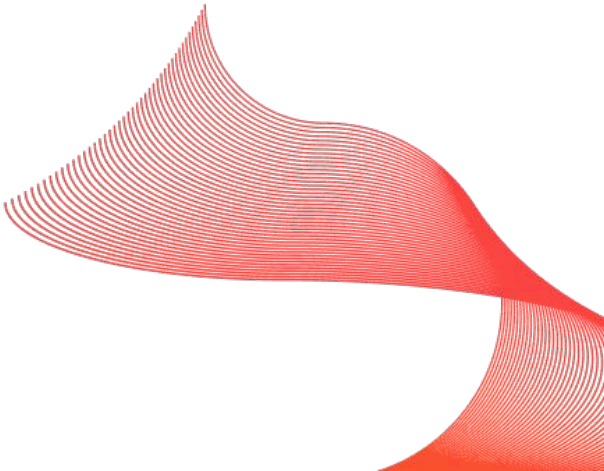
Description

A pair of headphones sitting on top of each other.

Name	Model
BLIP2	Fine-tuned Model

Description

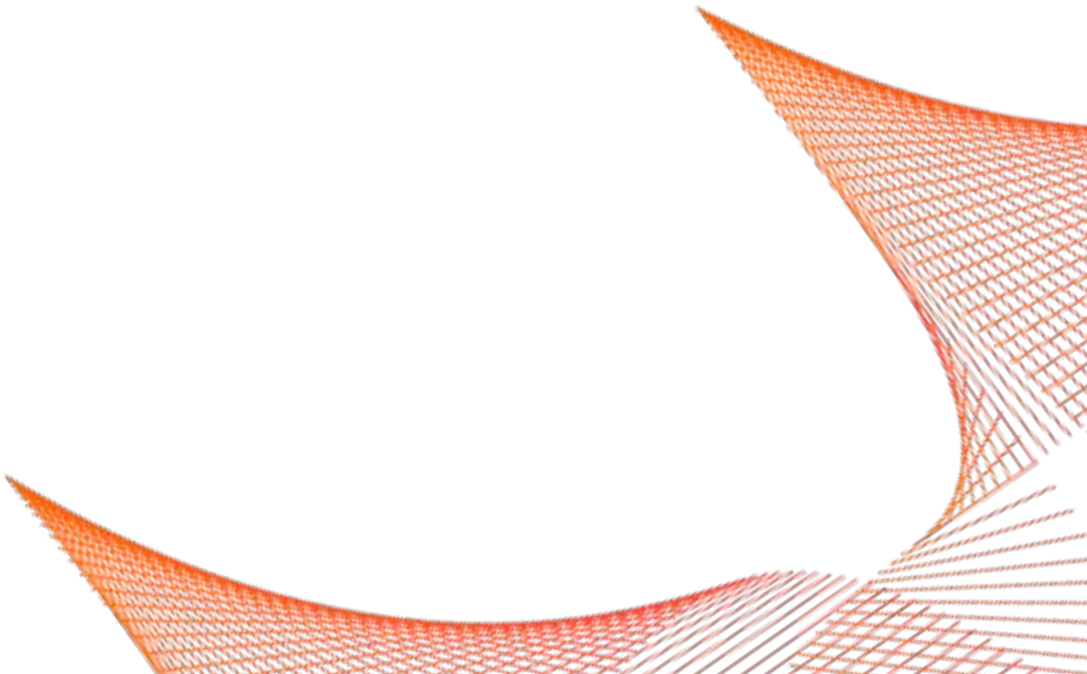
A pair of headphones taken from a close-up angle, highlighting the details of the headphones. The lighting is soft, highlighting the details of the headphones.



Conclusion

This study demonstrates that carefully curated, peer-ranked, and richly annotated datasets like the DSD significantly improve the performance and reliability of vision-language models. Fine-tuning experiments with LLAVA-NEXT and BLIP2 show enhanced syntactic precision, semantic fidelity, and stronger alignment between images and text-key factors for deep contextual understanding and human-like perceptual reasoning.

The DSD's multi-tier human-in-the-loop approach, combining perceptual judgments with technical annotations, validates a shift toward data-centric AI development. By providing a high-quality alternative to noisy crowd-sourced data, the DSD captures nuanced visual semantics and aesthetic preferences, aligning with current trends that prioritize data quality and context.



Though representing a fraction of the 100 million-plus images available through DataSeeds.AI, the DSD has broad potential as a foundational resource for AI tasks such as multimodal retrieval and generative image understanding. As AI models advance toward reasoning about complex real-world scenes, datasets like the DSD will be essential for pushing their capabilities further.

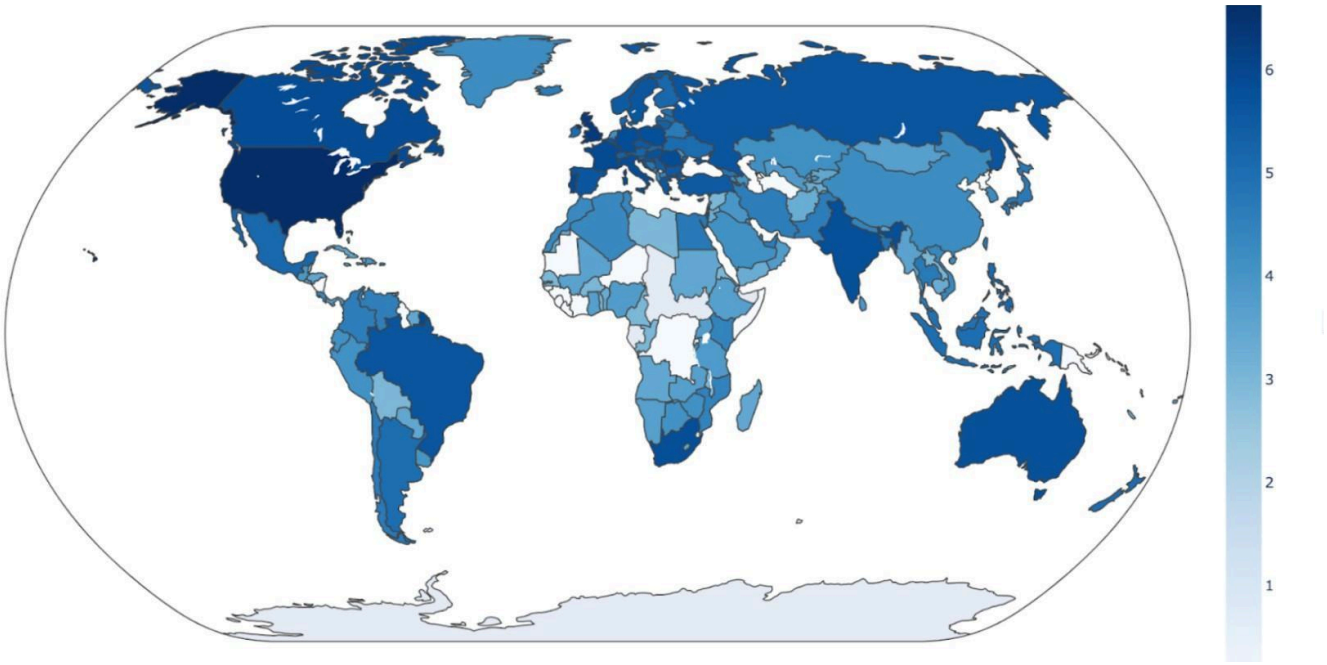


Figure 1: Gurushots Catalogue Geographic Distribution by Consented Image Count for AI Training (Log Scale)”

The DSD bridges human perceptual reasoning and machine inference, highlighting the critical role of high-quality, human-aligned data in advancing AI. We encourage the research community to adopt this resource in building more capable, context-aware, and reliable AI systems.

For more:

Research Paper

DSD Dataset

LLaVA weights

BLIP weights

Code scripts

Elevate Your AI

Perle

Partner with PerleAt Perle.

We accelerate AI development through expert data annotation, enrichment, and enhancement. Our platform combines a curated global expert network with AI-assisted workflows to deliver high-quality, multimodal datasets-designed to reduce rework and speed iteration.

Ready to elevate your AI training data? Contact us at sales@perle.ai or at perle.ai

DataSeeds.AI

Partner with DataSeeds.AI
DataSeeds.AI offers both on-demand and off-the-shelf image & video datasets enriched with detailed metadata, perfectly suited for AI model training. By leveraging a vast global network of creators and an extensive catalog, we provide rapid data collection and diverse content ensuring swift, scalable solutions to accelerate your AI projects.

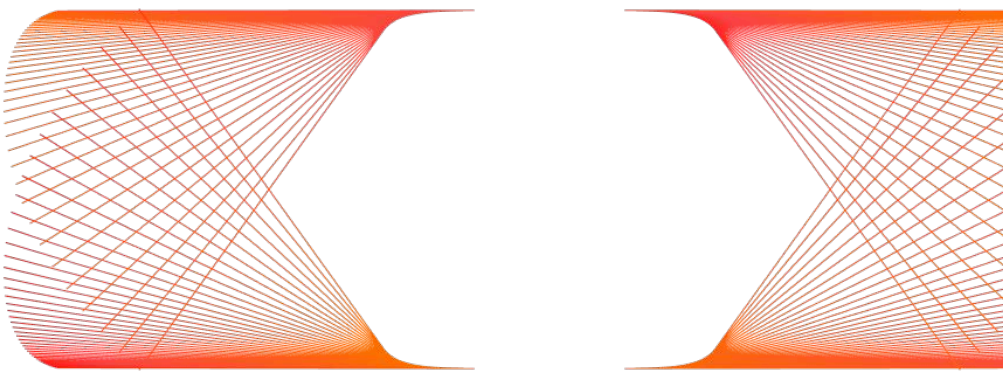
Unlock high-quality custom or ready-to-use datasets at DataSeeds.AI. Contact us at sales@dataseeds.ai

Brickroad

Partner with Brickroad

At Brickroad, we're solving AI's data availability challenges with our white glove supply and demand services that connect buyers and sellers instantly. Through deep industry partnerships and our native dataset marketplace, Brickroad, we offer sourcing, annotation, evaluation, research, compliance, licensing, liquidity, and sales solutions for suppliers and AI labs around the world.

Looking to optimize your data pipeline? Get in touch with Brickroad at freeman@brickroadapp.com



References

Nikita Bhatt, Nirav Bhatt, Purvi Prajapati, Vishal Sorathiya, Samah Alshathri, and Walid El-Shafai. A data-centric approach to improve performance of deep learning models. *Scientific Reports*, 14(1):22329, 2024.

Lian, Long, Yifan Ding, Yunhao Ge, Sifei Liu, Hanzi Mao, Boyi Li, Marco Pavone et al. "Describe anything: Detailed localized image and video captioning." arXiv preprint arXiv:2504.16072 (2025)