William Yang
4th year PhD
williamyang@princeton.edu

Ai Princeton Visual AI Lab

PRINCETON UNIVERSITY

Xindi Wu
3rd year PhD
xindiw@princeton.edu

Demystifying Computer Vision Systems

Data for Efficient Multimodal Machine Learning

ImageNet-OOD: Deciphering Modern Out-of-Distribution Detection Algorithms

William Yang*, Byron Zhang*, Olga Russakovsky   ICLR, 2024

To study impact of semantic shift without the influence of covariate shift in new-class detection, we construct **ImageNet-OOD** from ImageNet-21K
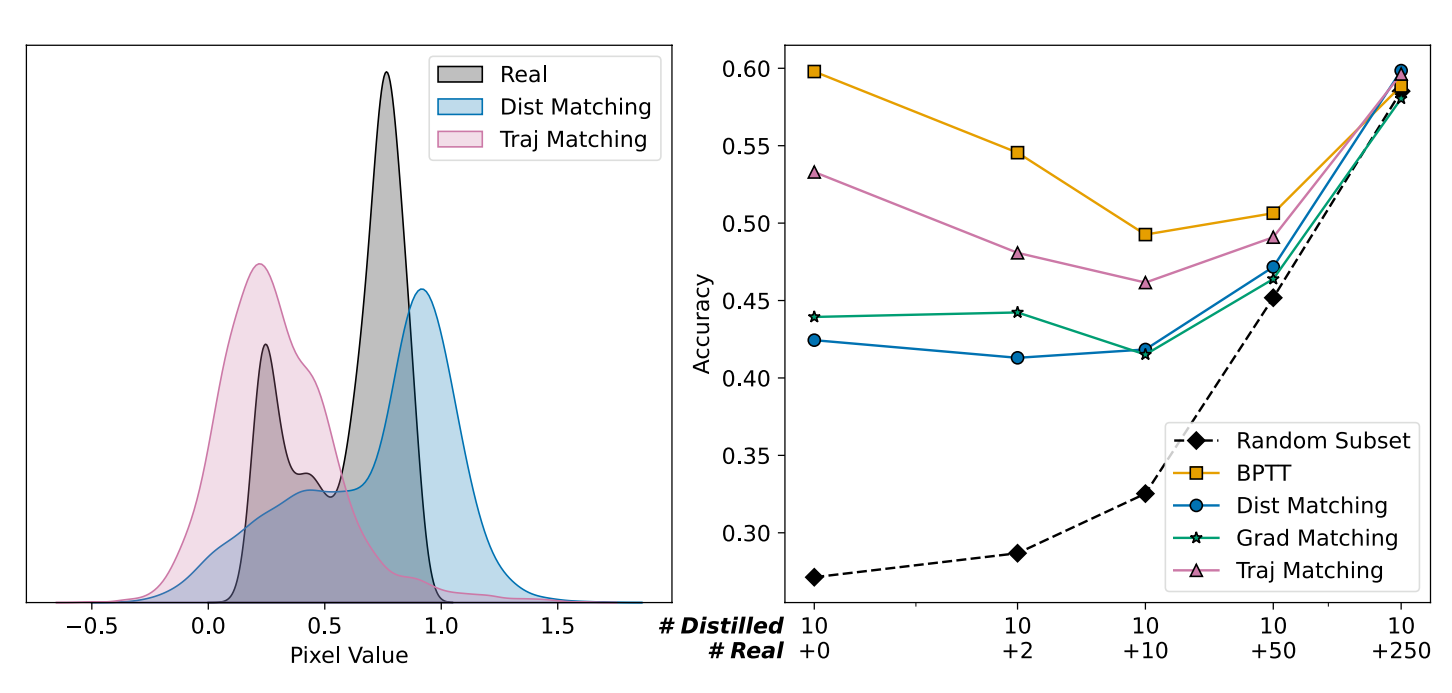
ImageNet-1K — Ice Cream | ImageNet-21K — Frozen Dessert

ImageNet-1K — Violin | ImageNet-21K — Viola

Semantic Ambiguity | Visual Ambiguity

ID Classes — Animal — Vehicle

OOD?

Class Ambiguity w.r.t the Decision Boundary

What are modern OOD detectors detecting?

One qualitative example on how covariate shift affects OOD detection performance

Original Image | Zoomed Image

Ostrich: 1.00 | **Ostrich: 0.98**

ImageNet-OOD Percentile by Method

Method		
MSP	100.0	95.3
Energy	100.0	83.7
ViM	100.0	61.6
ReAct	99.9	60.4

Real data | **Distilled data**

BPTT — subset, early stopping
Distribution Matching
Gradient Matching
Trajectory Matching

Predictions from models trained on distilled data are more similar to models that were early-stopped on the full data.

What is Dataset Distillation Learning

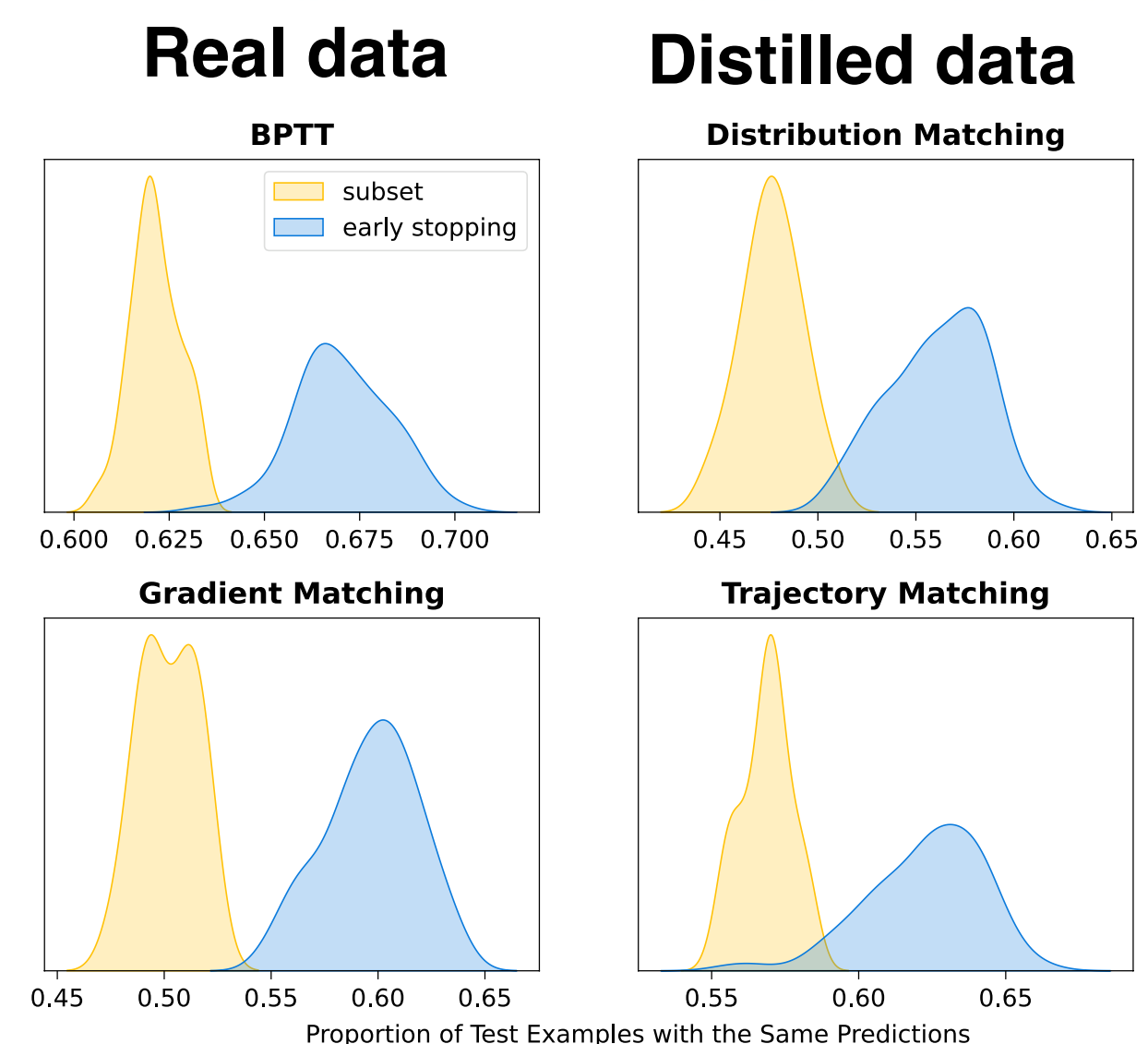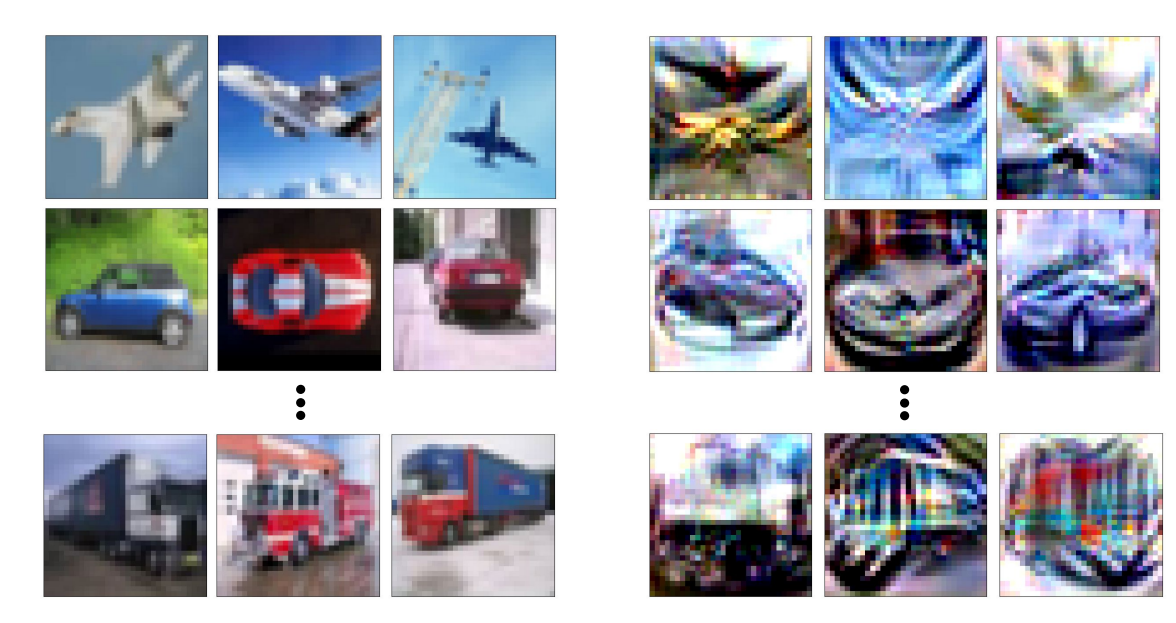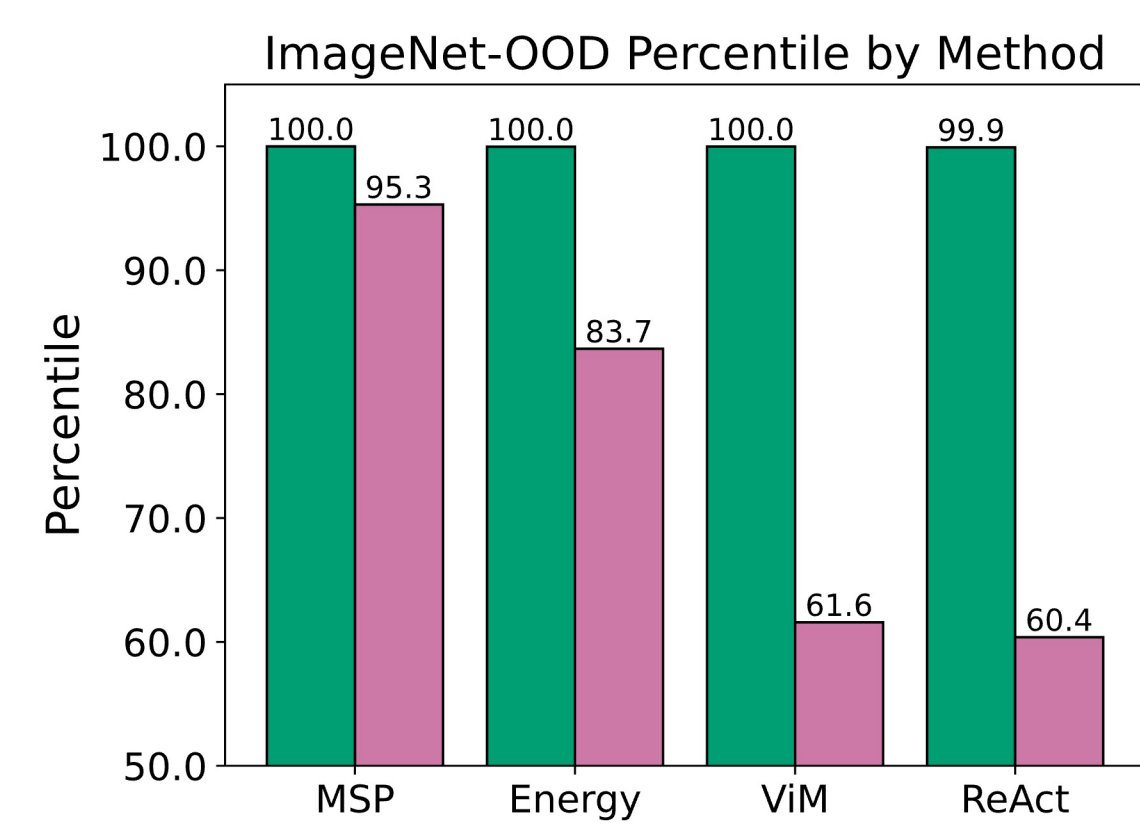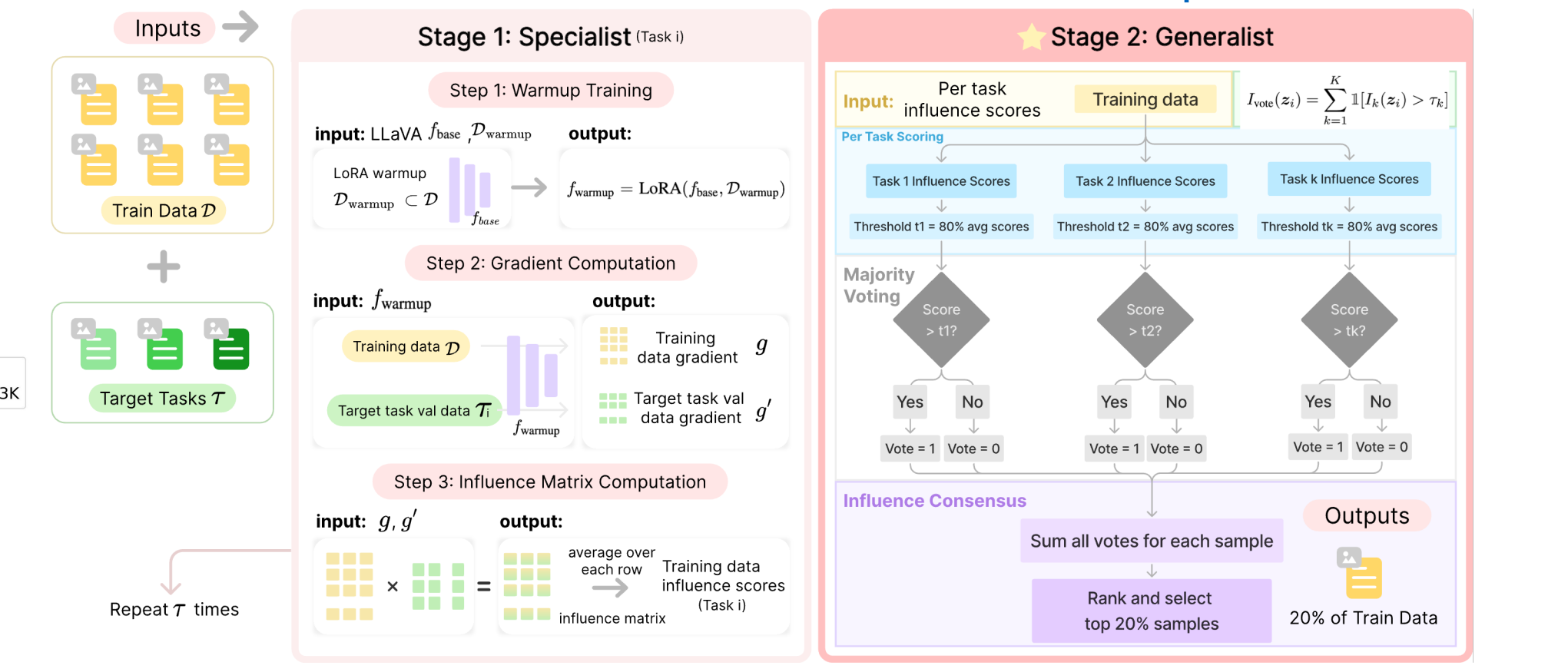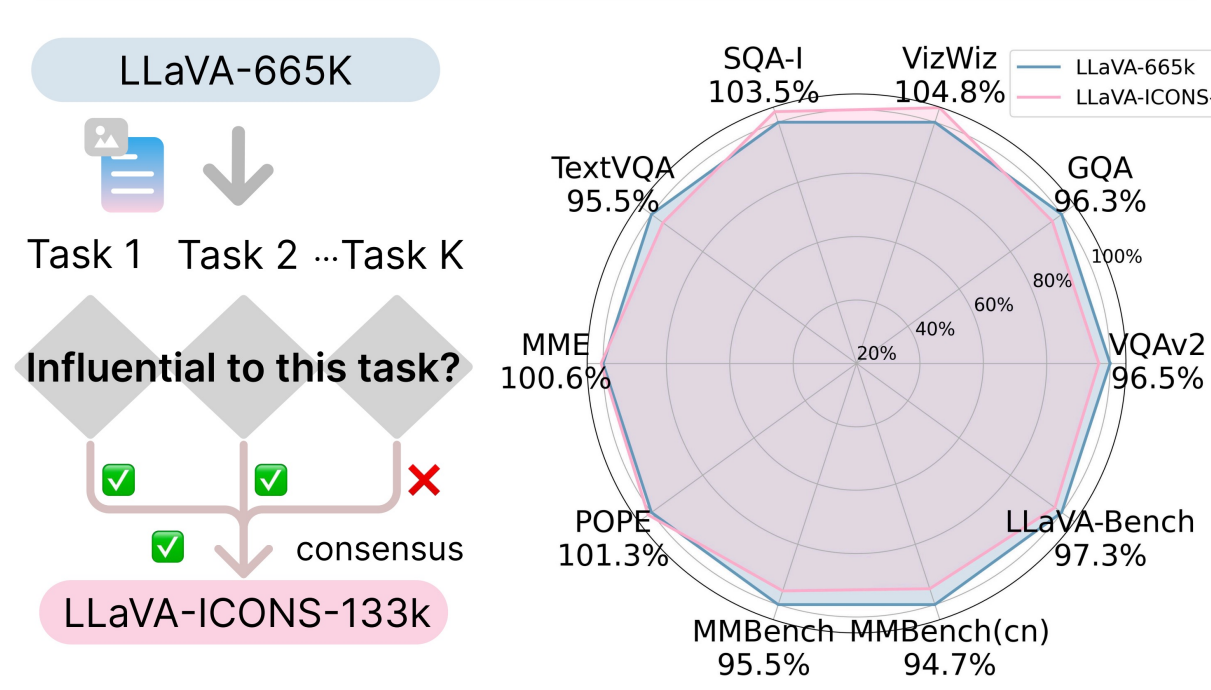William Yang, Ye Zhu, Zhiwei Deng, Olga Russakovsky   ICML, 2024

Dataset distillation learns a compact set of synthetic data that retains essential information from the original dataset, but why this is possible and what do they represent remains unclear.

Real, Dist Matching, Traj Matching

Random Subset, BPTT, Dist Matching, Grad Matching, Traj Matching

The distilled data lies outside of the real data manifold and is sensitive during training.

ICONS: Influence Consensus for Vision-Language Data Selection

Xindi Wu, Mengzhou Xia, Rulin Shao, Zhiwei Deng, Pang Wei Koh, Olga Russakovsky

Arxiv 2025   Website Paper Code Dataset

Can we identify a compact subset of training data that preserves model capabilities while enabling faster experimentation?

LLaVA-665K → Task 1 Task 2 ··· Task K

Influential to this task? ✅ ✅ ❌

✅ consensus

LLaVA-ICONS-133k

Stage 1: Specialist (Task i)
Step 1: Warmup Training
input: LLaVA f_{base} D_{warmup} → output: $f_{warmup} = $ LoRA(f_{base}, D_{warmup})
LoRA warmup $D_{warmup} \subset D$
Step 2: Gradient Computation
input: f_{warmup} → output: Training data gradient g
Training data D
Target task val data T → Target task val data gradient g'
Step 3: Influence Matrix Computation
input: g, g' → output: average over each row — Training data influence scores (Task i)
Repeat T times

Stage 2: Generalist
Input: Per task influence scores, Training data
$I_{cons}(z_i) = \sum_{k=1}^{K} \mathbb{1}[I_k(z_i) > \tau_i]$
Per Task Scoring — Task 1 Influence Scores, Task 2 Influence Scores, Task k Influence Scores
Threshold t1 ≥ 80% avg scores, Threshold t2 ≥ 80% avg scores, Threshold tk ≥ 80% avg scores
Majority Voting — Score > t1, Score > t2, Score > tk — Yes Vote = 1 / No Vote = 0
Influence Consensus — Sum all votes for each sample
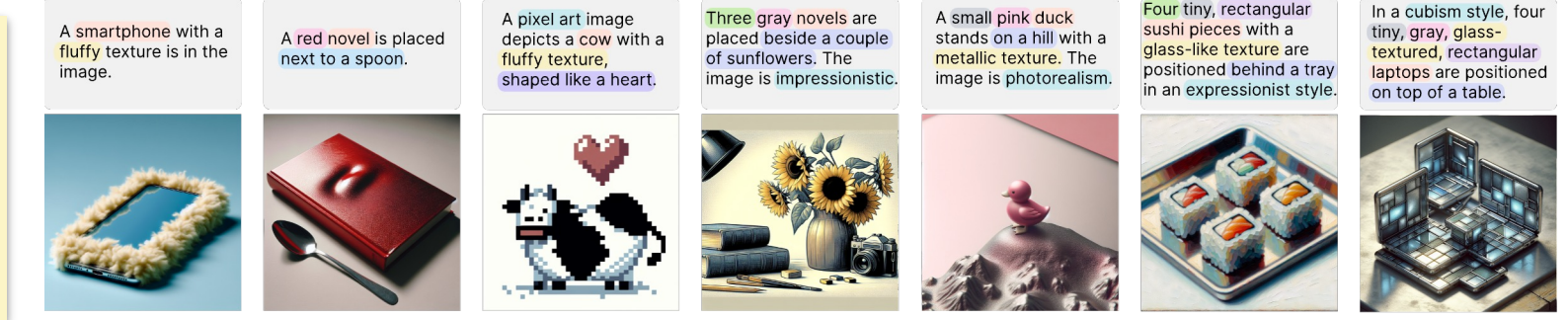Outputs — Rank and select top 20% samples — 20% of Train Data

Key results:
A 20% subset of LLaVA-665K and a 20% subset of Cambrian-7M achieve 98.6% and 98.8% of the rel. obtained using the full datasets. LLaVA 60% subset achieves >100% rel. Demonstrated strong transferability across unseen tasks and scalability to larger models.

Radar chart: SQA-I 103.5%, VizWiz 104.8%, GQA 96.3%, VQAv2 96.5%, LLaVA-Bench 97.3%, MMBench(cn) 94.7%, MMBench 95.5%, POPE 101.3%, MME 100.6%, TextVQA 95.5% — LLaVA-665k, LLaVA-ICONS-133K
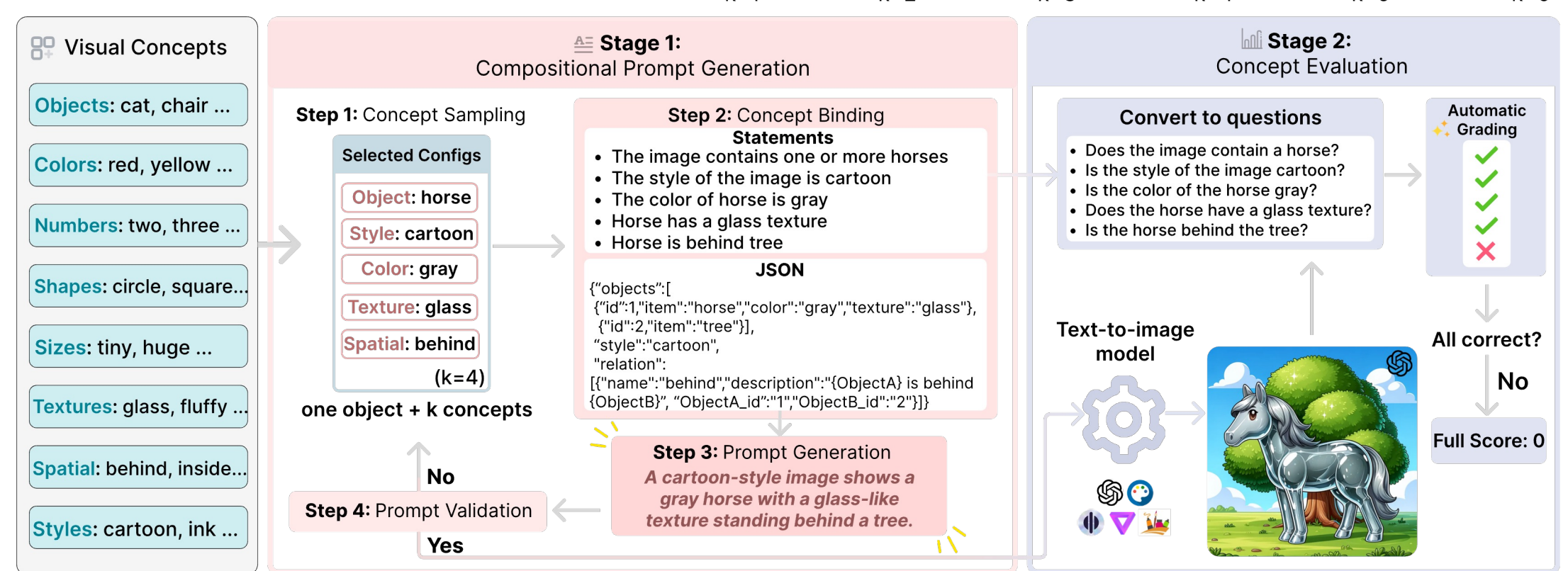
ConceptMix: A Compositional Image Generation Benchmark with Controllable Difficulty

Xindi Wu*, Dingli Yu*, Yangsibo Huang*, Olga Russakovsky, Sanjeev Arora

NeurIPS, 2024

Website Paper Code

How to automatically generate prompts that mix diverse visual concepts?

Concept categories: object, texture, color, shape, style, number, spatial, size

k = # visual concepts — k=1, k=2, k=3, k=4, k=5, k=6, k=7

How to automatically grade results based on (prompt, generated image)?

Visual Concepts
Objects: cat, chair ...
Colors: red, yellow ...
Numbers: two, three ...
Shapes: circle, square...
Sizes: tiny, huge ...
Textures: glass, fluffy ...
Spatial: behind, inside...
Styles: cartoon, ink ...

Stage 1: Compositional Prompt Generation
Step 1: Concept Sampling — Selected Configs: Object: horse, Style: cartoon, Color: gray, Texture: glass, Spatial: behind (k=4) — one object + k concepts
Step 2: Concept Binding
Statements
- The image contains one or more horses
- The style of the image is cartoon
- The color of the horse is gray
- Horse has a glass texture
- Horse is behind the tree
JSON
"objects":[
{"id":1,"item":"horse","color":"gray","texture":"glass"},
{"id":2,"item":"tree"}],
"style":"cartoon",
"relation":
[{"name":"behind","description":"(ObjectA) is behind (ObjectB)", "ObjectA_id":"1","ObjectB_id":"2"}]
Step 3: Prompt Generation
A cartoon-style image shows a gray horse with a glass-like texture standing behind a tree.
Step 4: Prompt Validation — No / Yes

Stage 2: Concept Evaluation
Convert to questions:
- Does the image contain a horse?
- Is the style of the image cartoon?
- Is the color of the horse gray?
- Does the horse have a glass texture?
- Is the horse behind the tree?
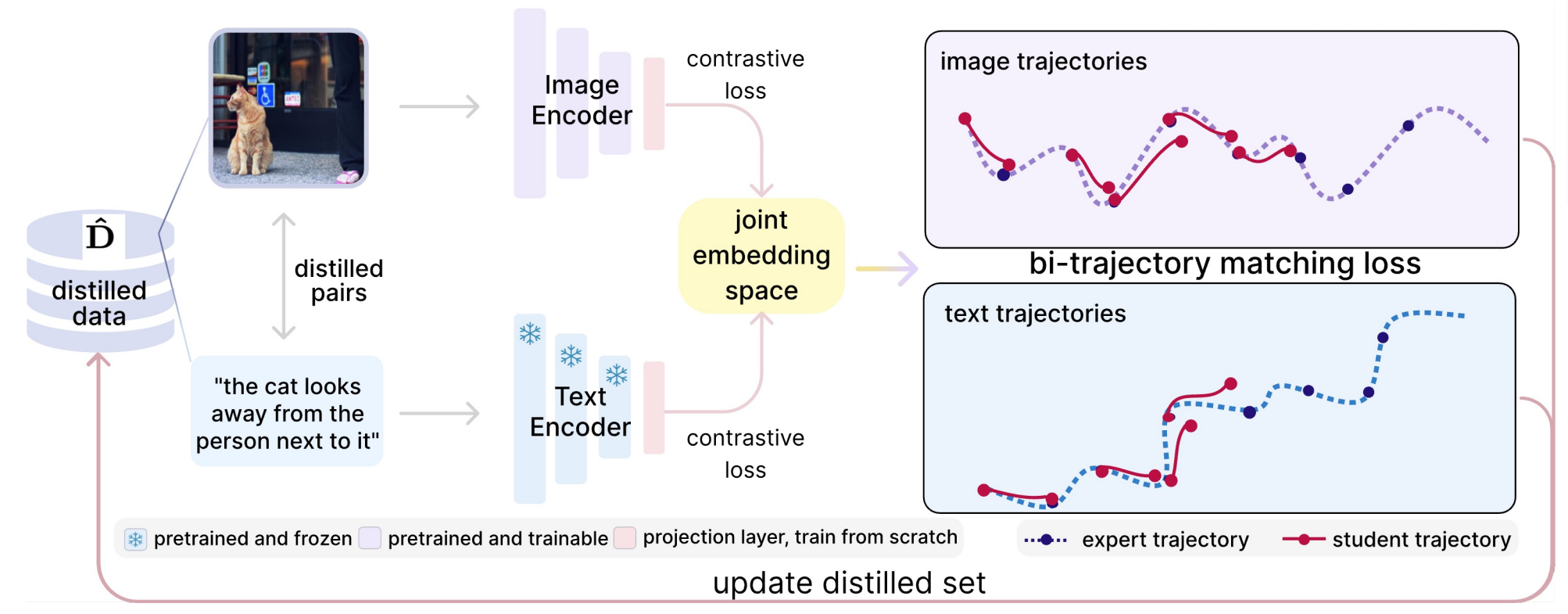Automatic Grading ✅✅❌ — All correct? No — Full Score: 0
Text-to-image model

Key results:
Color and style are the easiest while spatial, size, and shape are challenging.
We observe a consistent performance drop as K increases with the leading model DALLE 3, struggles at k=5.

Why are models bad at $k \geq 3$?
- Limited exposure to complex concept combinations.
- Disparate concept representation.

Vision-Language Dataset Distillation

Xindi Wu, Byron Zhang, Zhiwei Deng, Olga Russakovsky

How can we distill the most critical information from vision-language datasets?

Website arXiv Code   TMLR, 2024

$\hat{D}$ distilled data — distilled pairs — "the cat looks away from the person next to it"
Image Encoder → contrastive loss → image trajectories
Text Encoder → contrastive loss → text trajectories
joint embedding space → bi-trajectory matching loss
pretrained and frozen | pretrained and trainable | projection layer, train from scratch
expert trajectory | student trajectory — update distilled set

• **Bi-trajectory matching**: Separately considers two trajectories to capture complex vision-text interactions via contrastive loss.

Distilled examples:
a newly married couple sharing a kiss in front of a convertible | a couple kissed in front of a beautiful three-tiered cake with blue ribbon and pink accents | a man in a black suite is surfing a huge wave in the beautiful blue water | a man surfs over a huge wave in the ocean

Key results:
Our method doubles the performance of coreset selection approaches with an order of magnitude fewer distilled pairs.