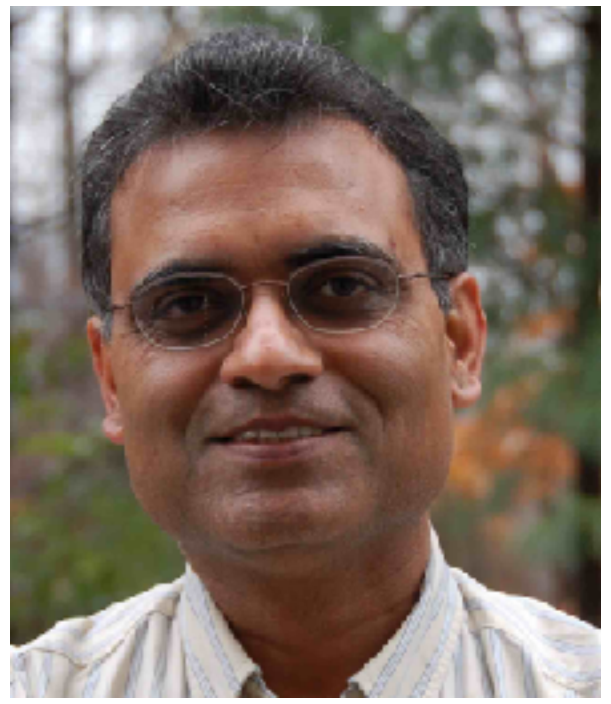
Test-Time Canonicalization by Foundation Models for Robust Perception

Utkarsh Singhal*

Ryan Feng*

Stella X. Yu

Atul Prakash


* denotes equal contribution

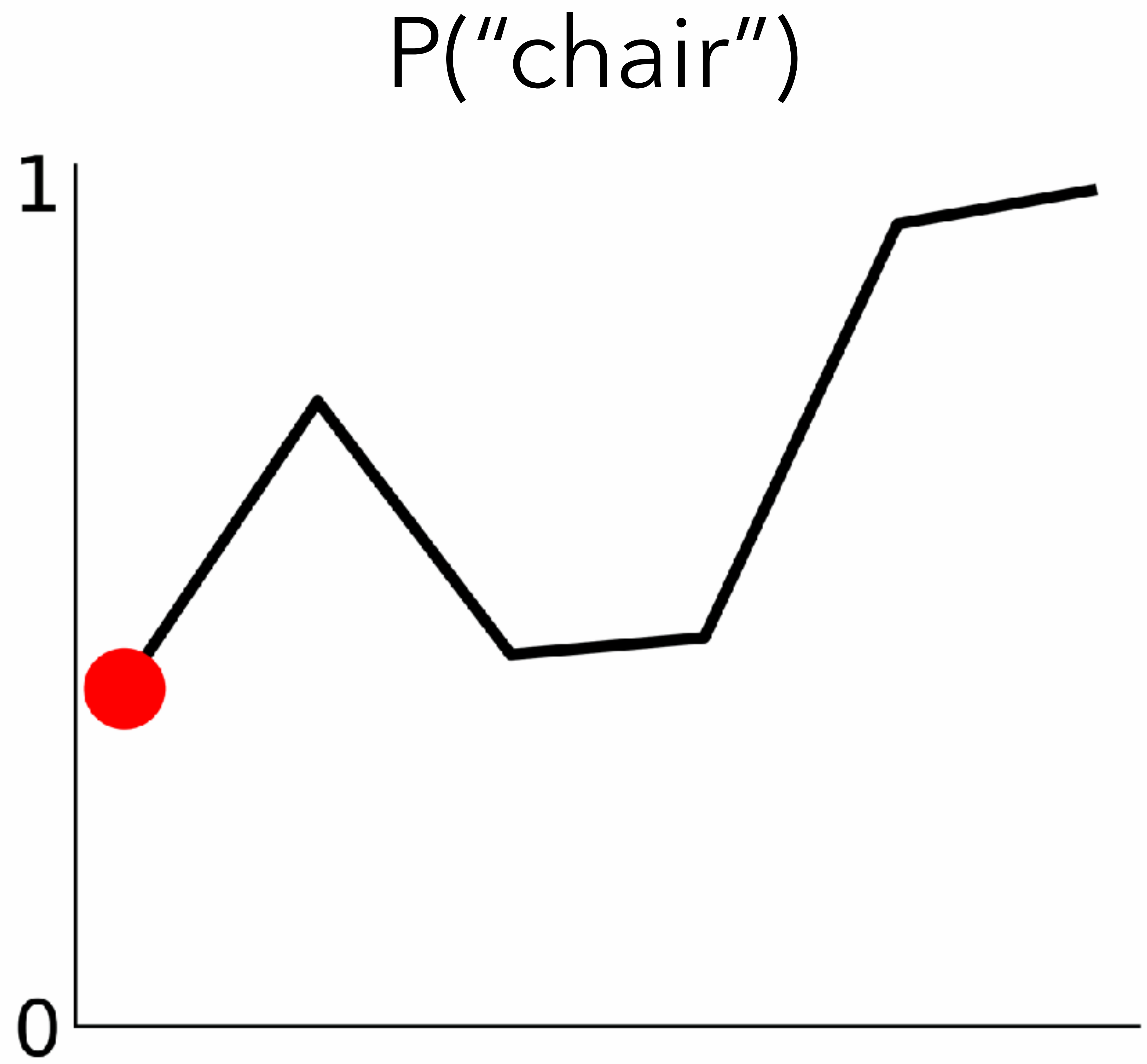
Mobile Agents Face Difficult and Diverse Input Transformations

Viewpoint

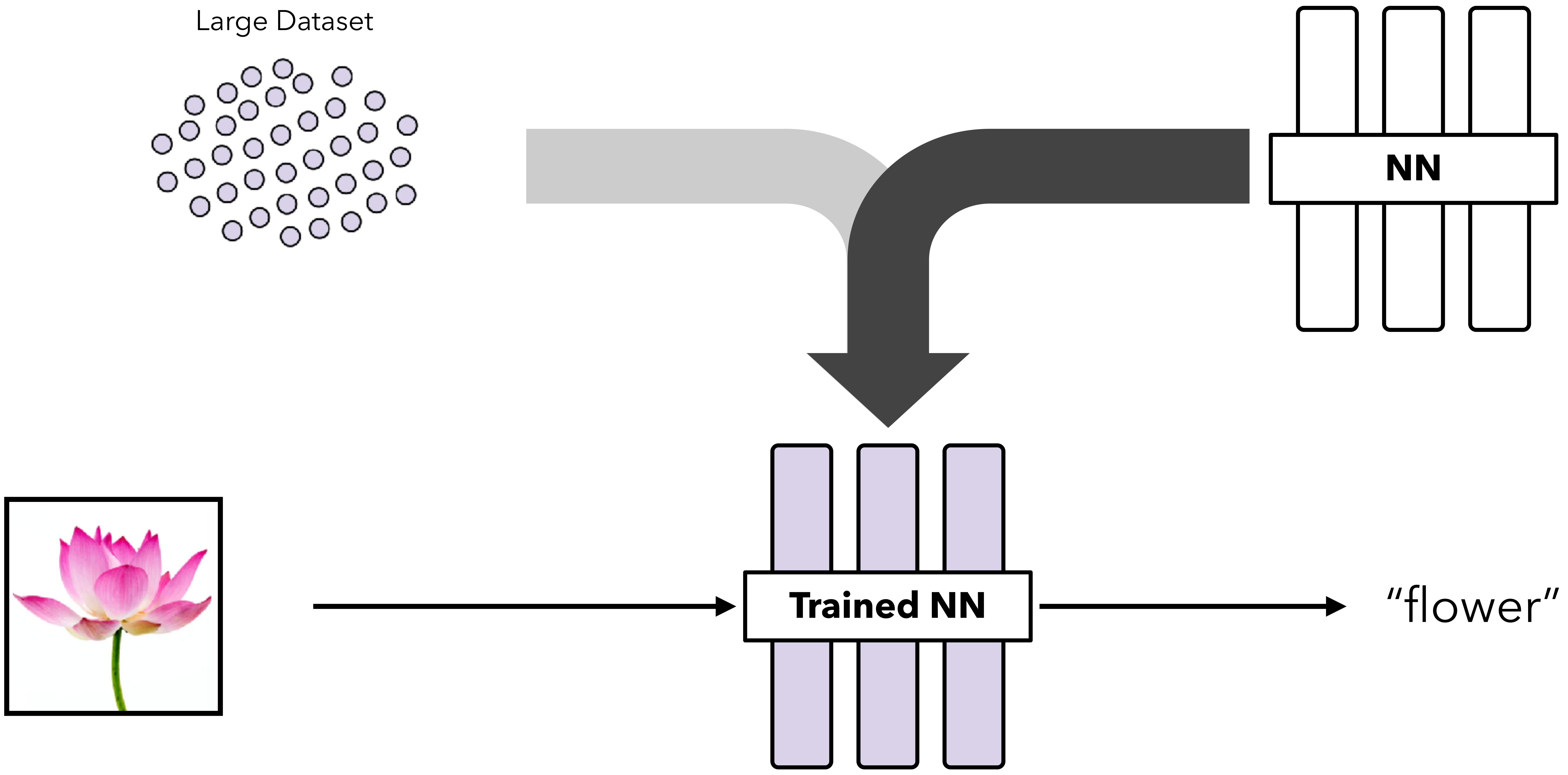
Lighting

Environment

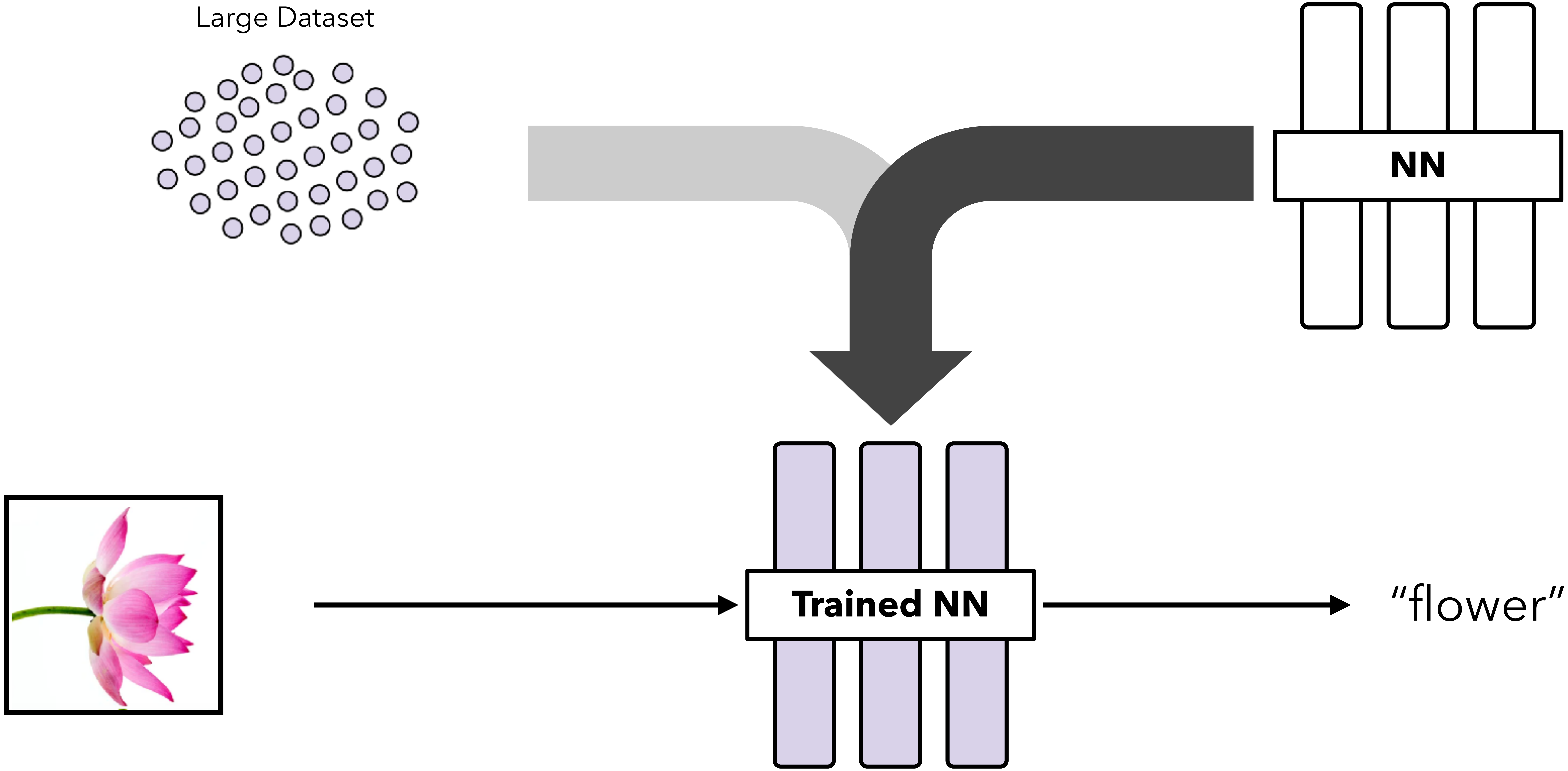
Foundation Models are Still not Robust Enough



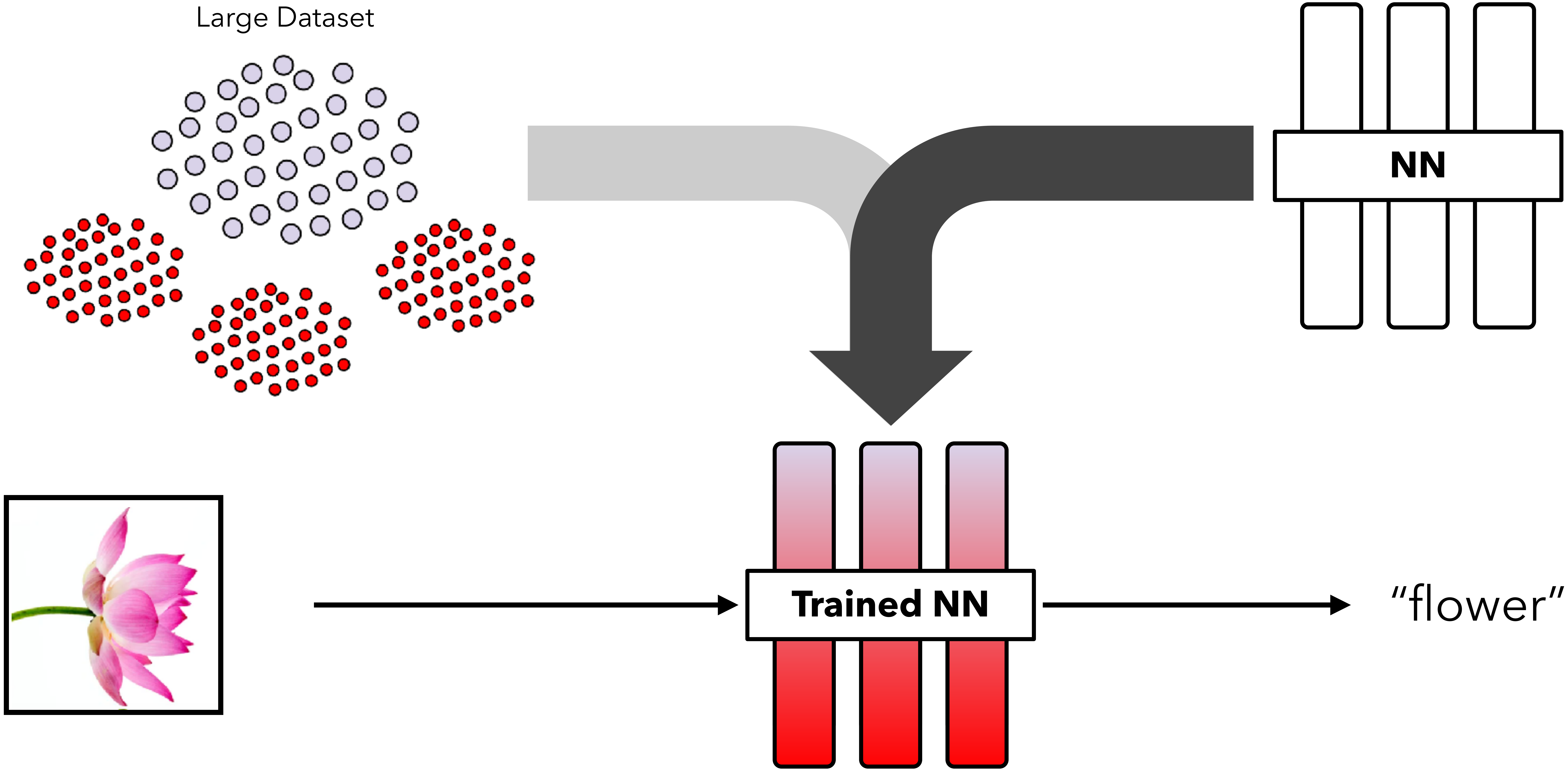
A Standard Pipeline Handles Upright Data



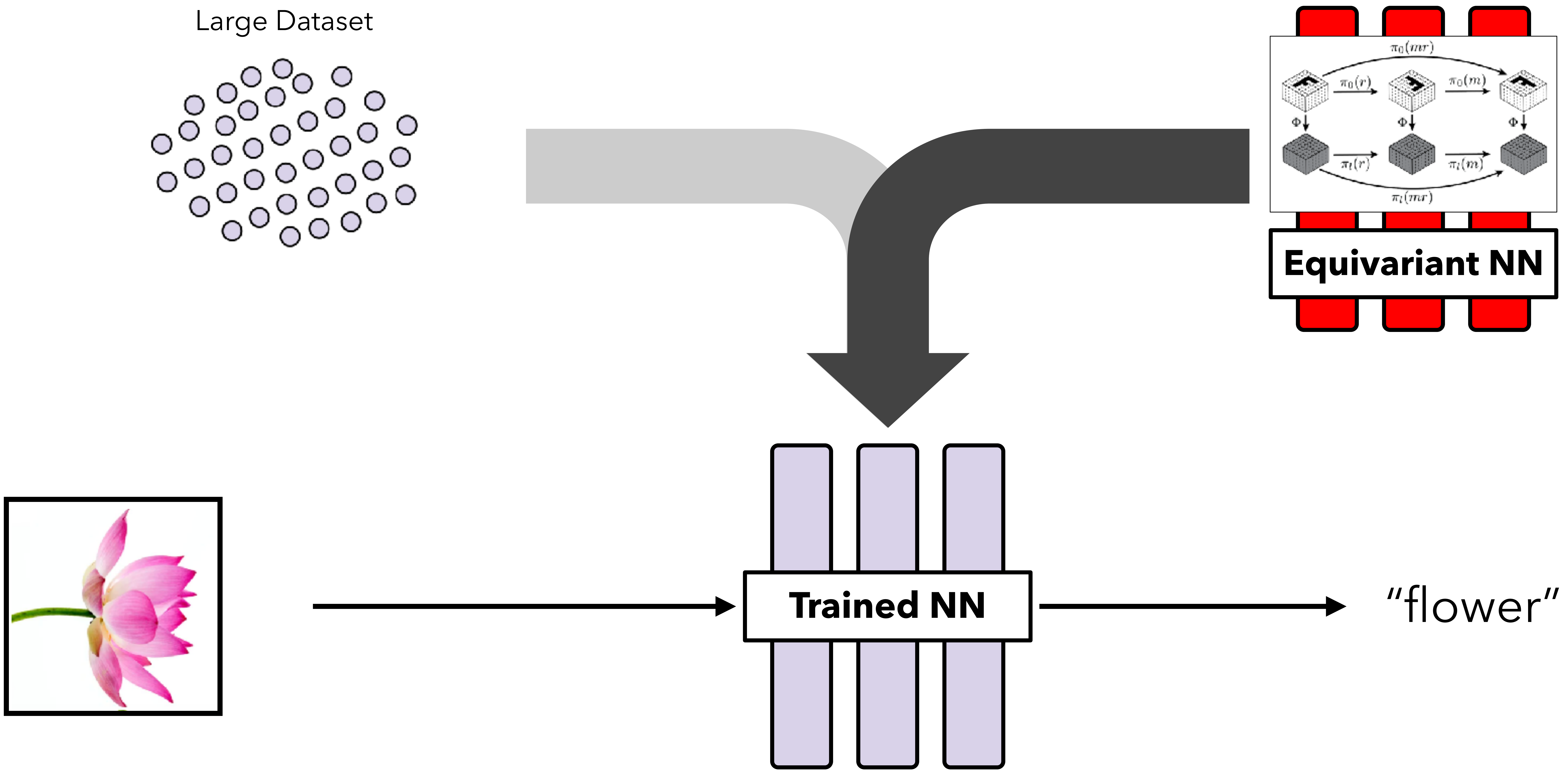
How to Handle Transformed Inputs?



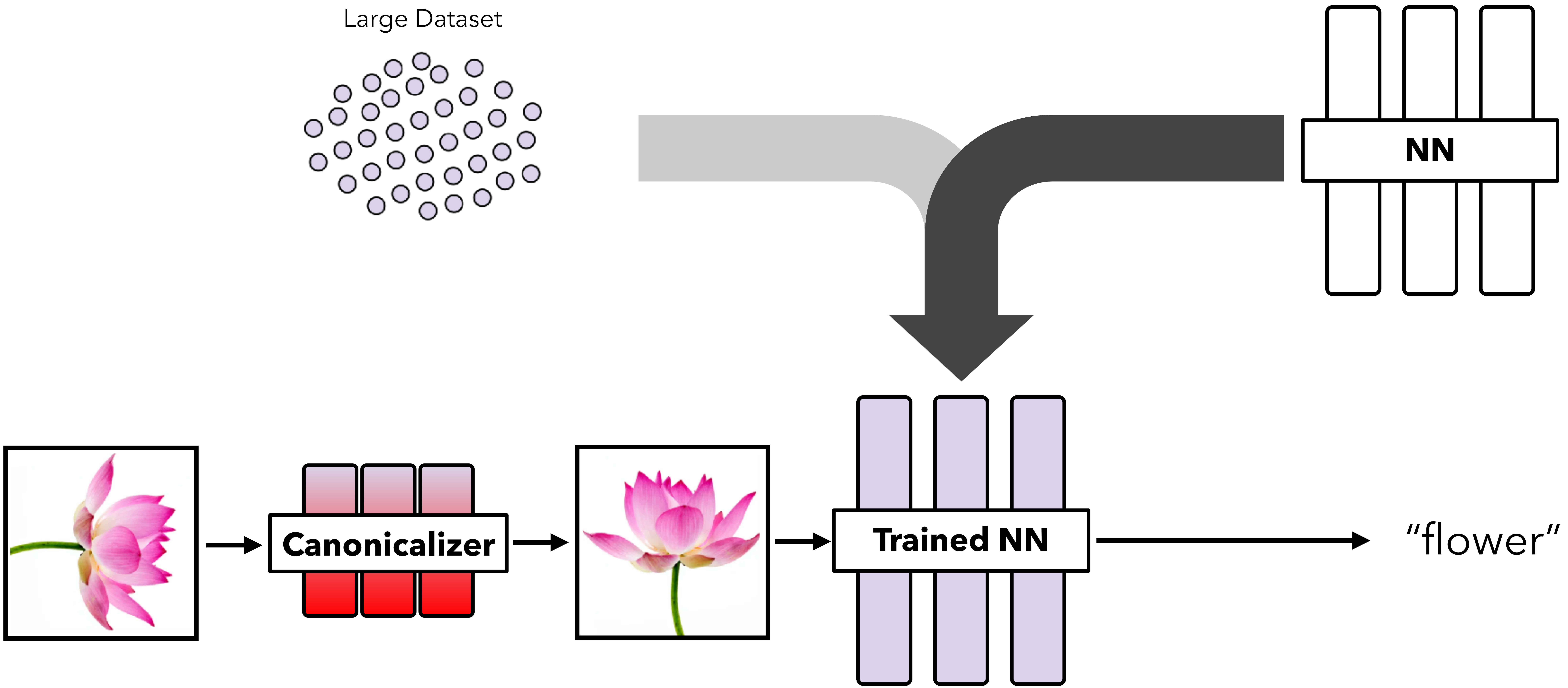
Data Augmentation: Train on Transformed Data



Equivariant Networks: Transform-Specific Architectures



Prior Canonicalization Work: Train a Network to Fix the Input



How to Help Foundation Models Face Diverse and Challenging Input Transformations?

- ~~Data augmentation?~~
  Train-time transforms only

- ~~Equivariant Neural Networks?~~
  Architecture-specific transforms only

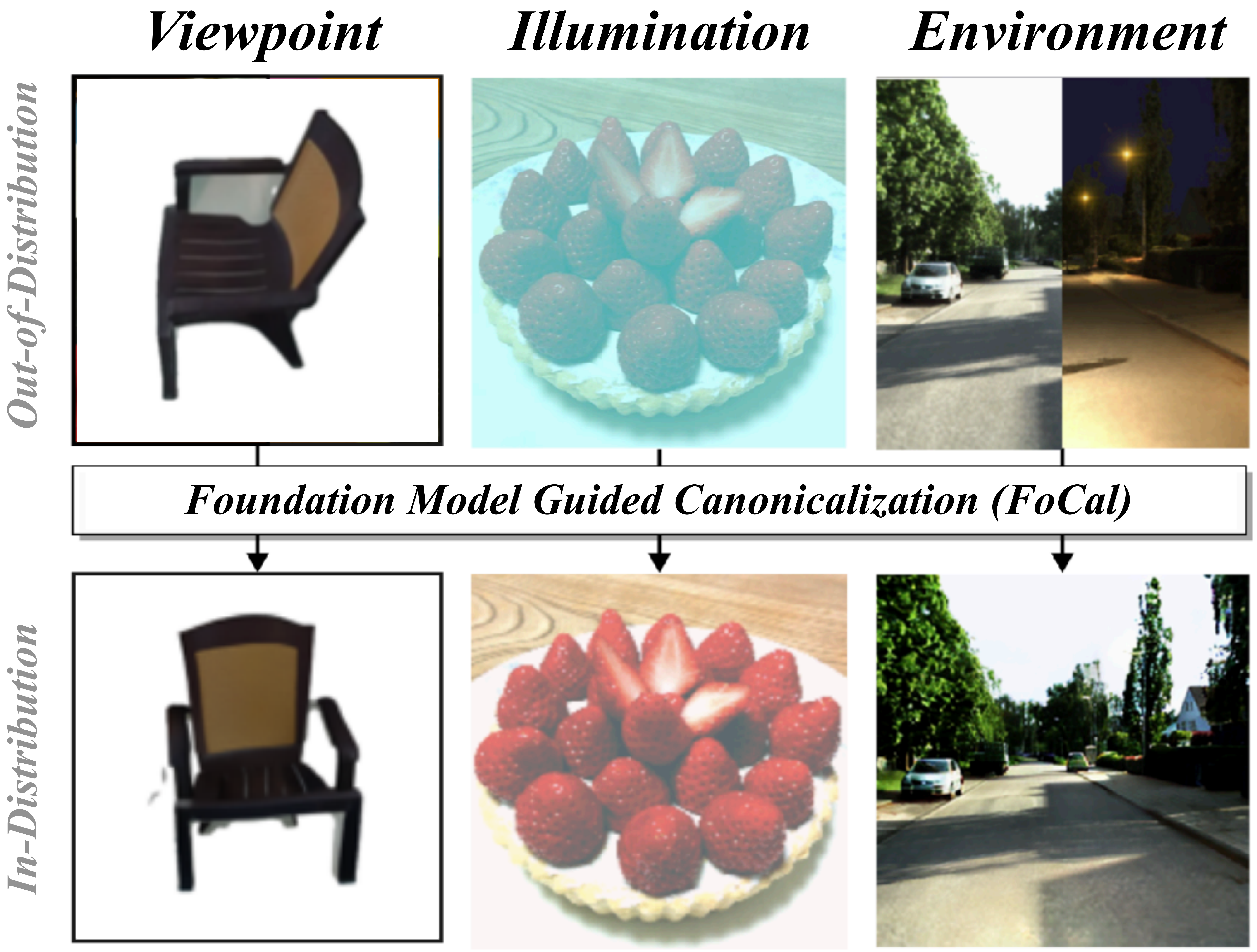
- ~~Prior Canonicalization work?~~
  Don’t generalize to new datasets & transforms

FoCal: Foundation Model Guided Canonicalization

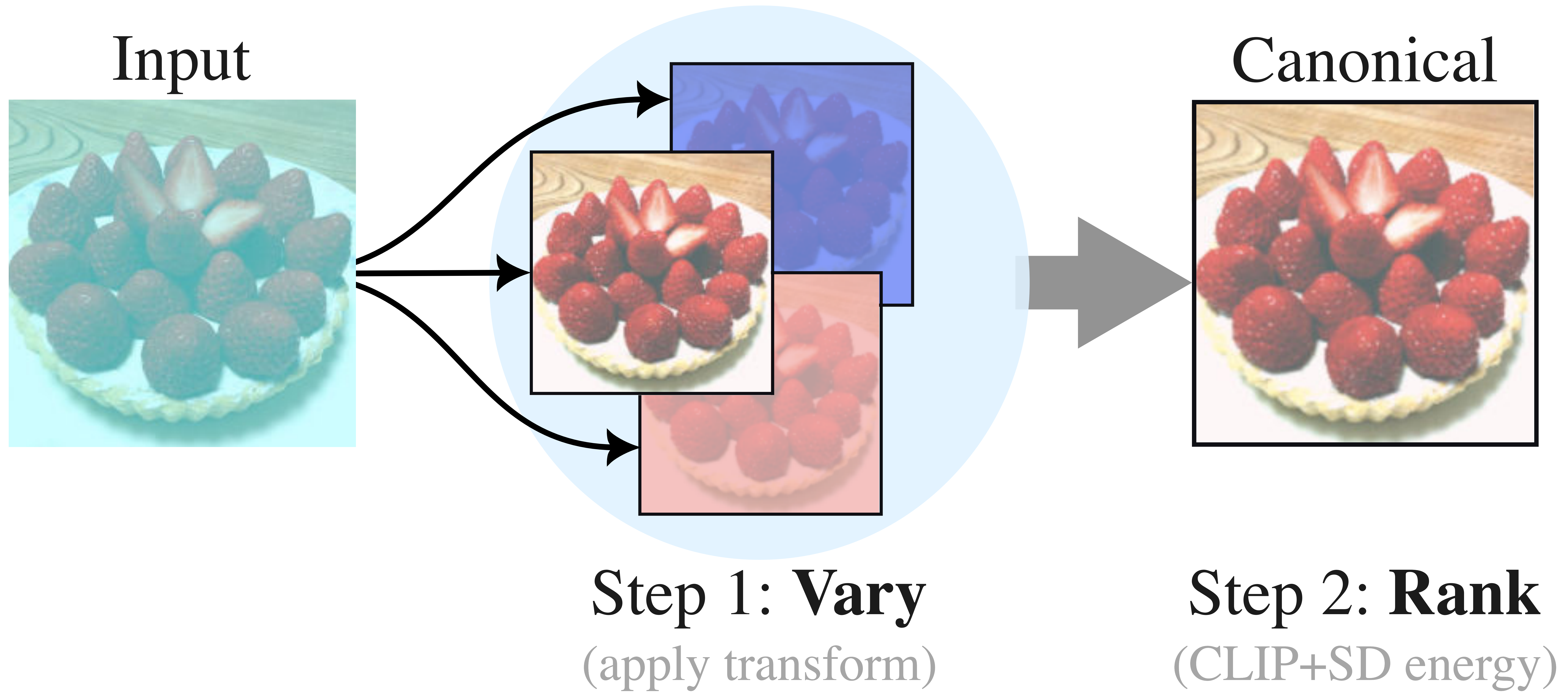
Insight: Foundation models know what is typical; we use this to convert OOD inputs ⟶ typical

Benefits:

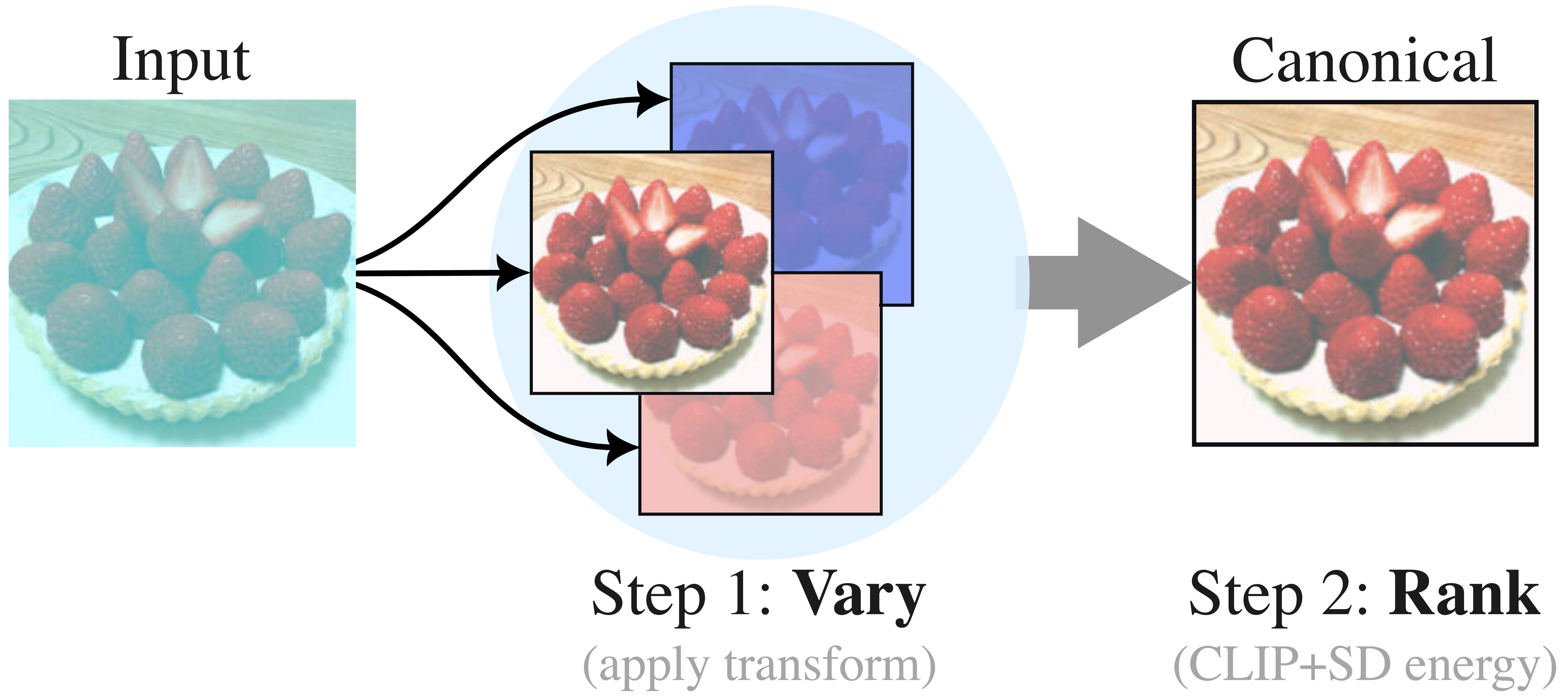
- Fully test-time approach
- Generalizes to diverse and complex transformations



FoCal: Test-Time Canonicalization by Foundation Models



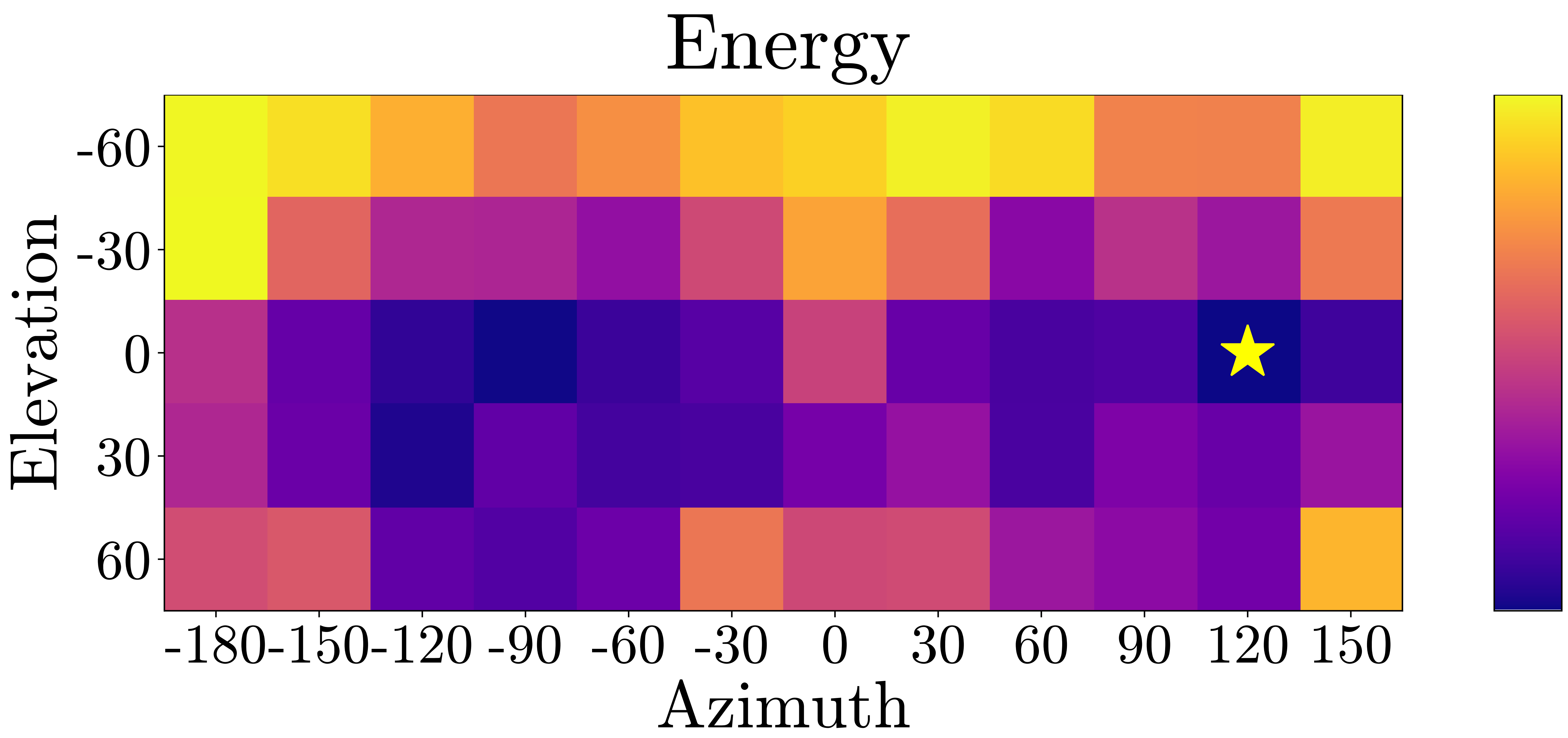
FoCal: Test-Time Canonicalization by Foundation Models



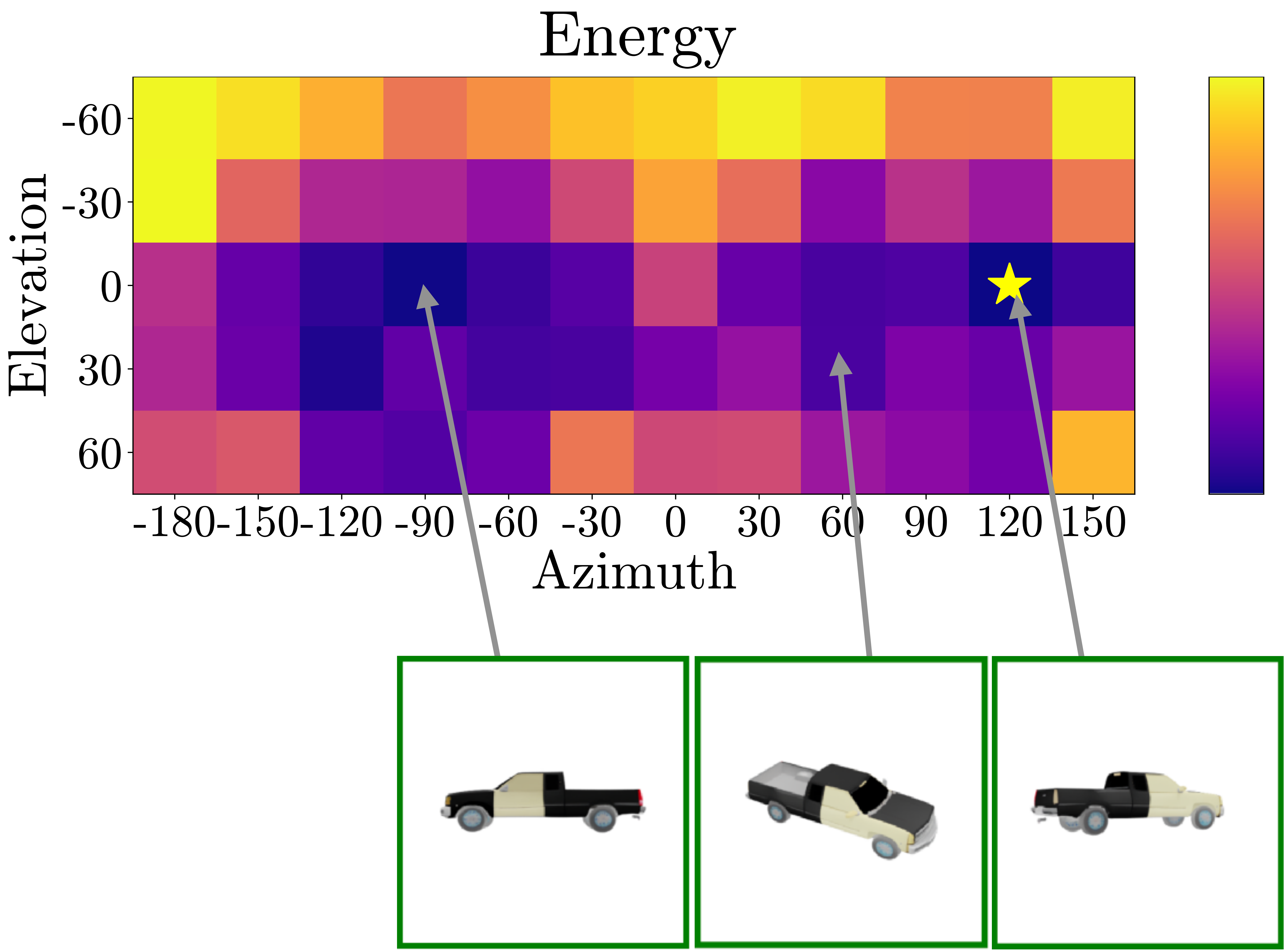
FoCal searches over the transformation space and selects the best version:

1. **Vary**: Generate candidate transformed versions
2. **Rank**: Find the best using foundation model priors (energy functions)

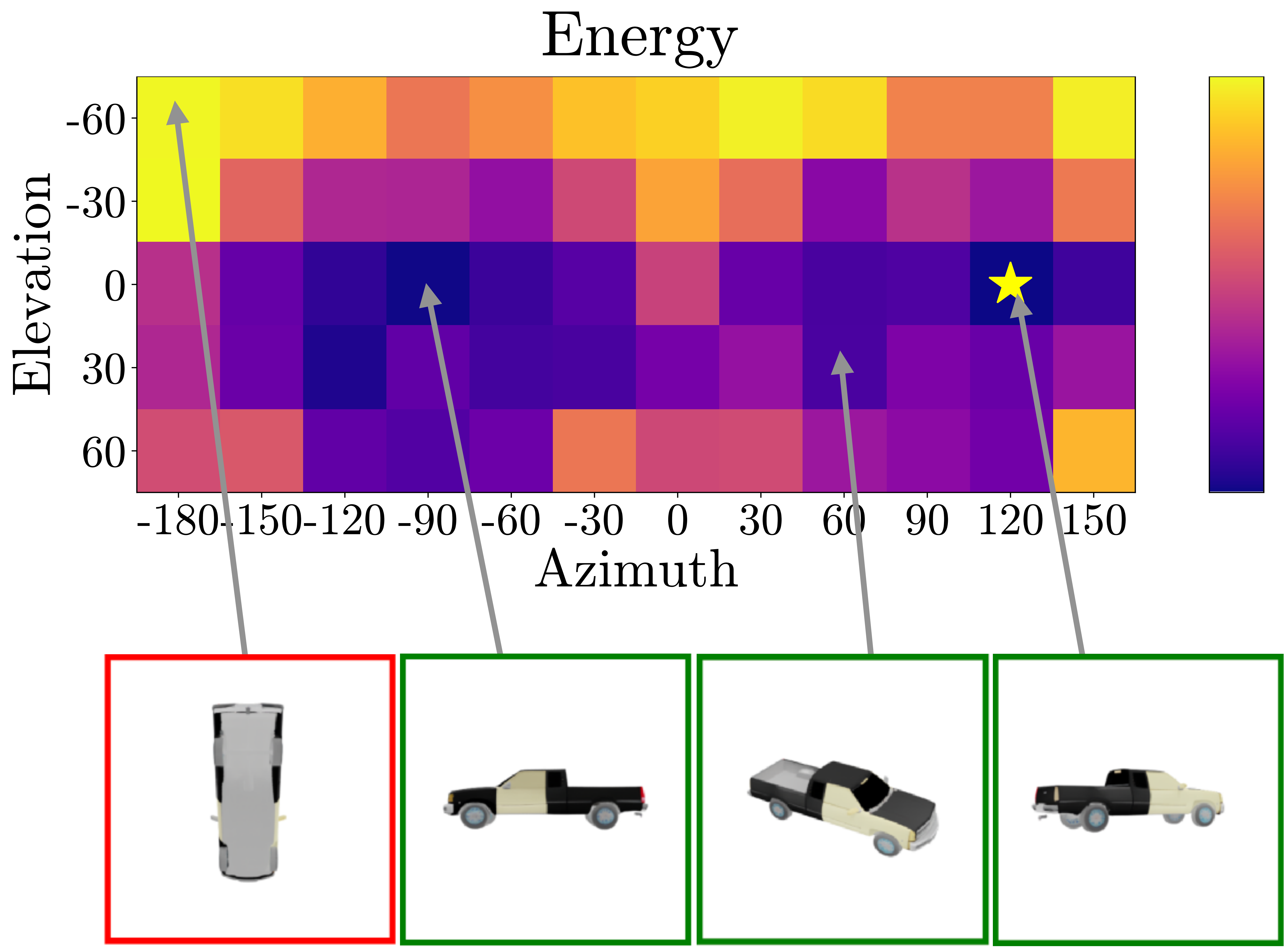
FoCal Energy Function Picks Typical Viewpoints



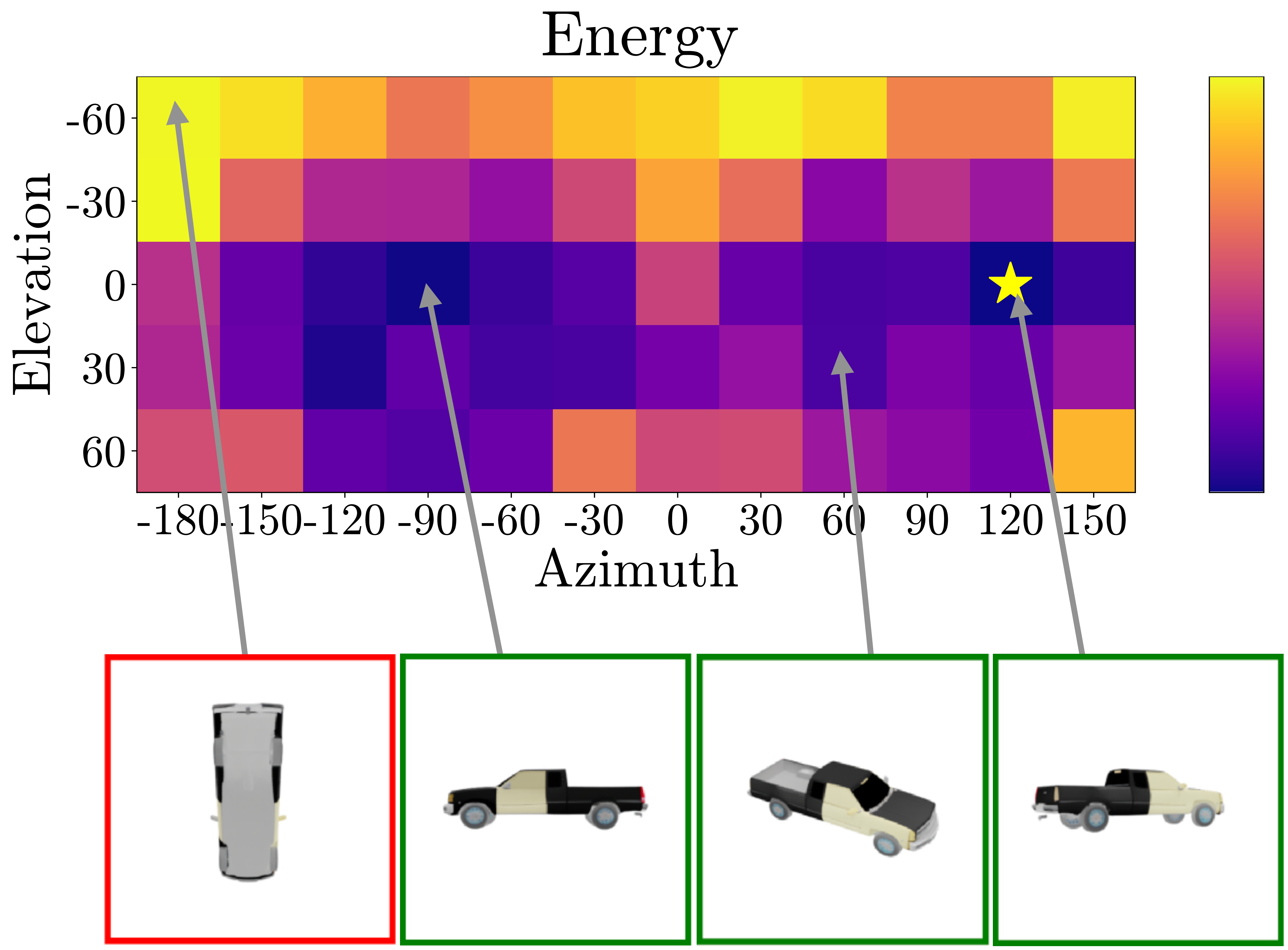
FoCal Energy Function Picks Typical Viewpoints



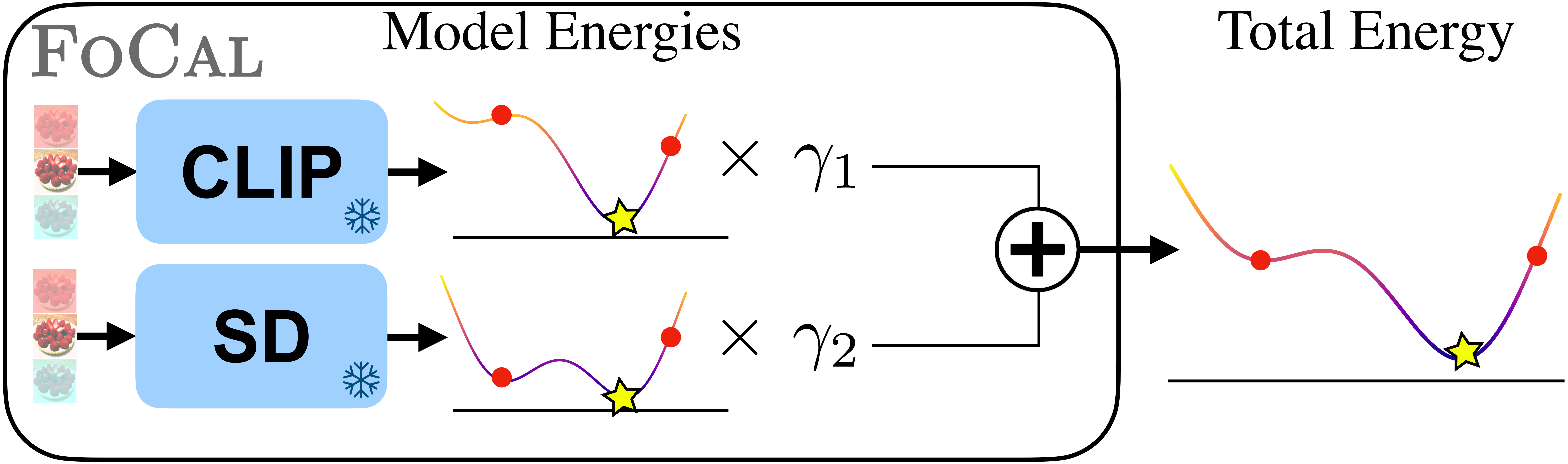
FoCal Energy Function Picks Typical Viewpoints



FoCal Energy Function Picks Typical Viewpoints



FoCal Energy = CLIP Prior + Stable Diffusion Prior



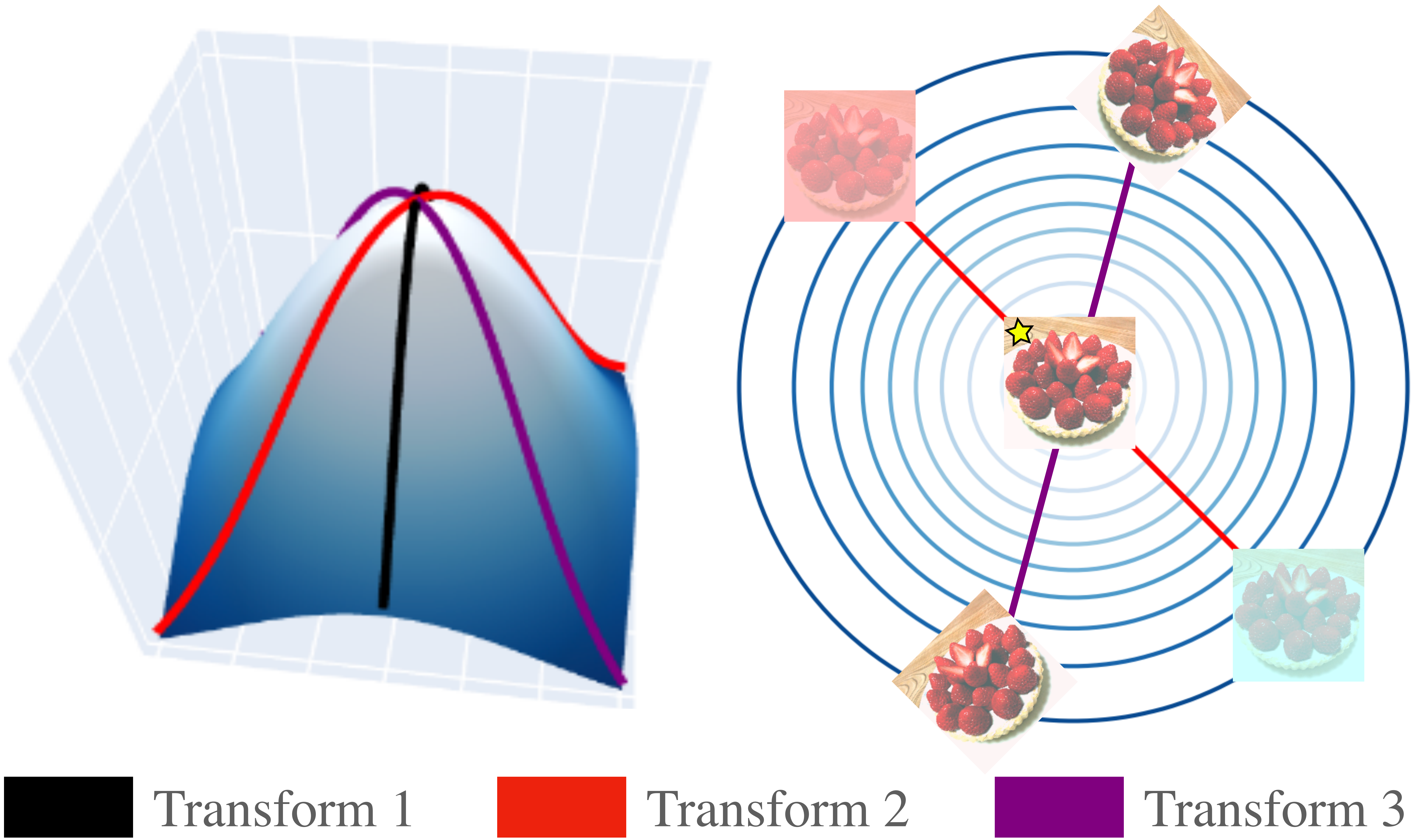
$$E_{\text{CLIP}}(\boldsymbol{x};\alpha,\beta) = \Big(\alpha\cdot \underset{c\in 1,2,\ldots,|C|}{\text{mean}} - \beta\cdot\max\Big)\big(f_\theta(\boldsymbol{x})[c]\big)$$

Classifier Energy

$$E_{\text{diff}}(\boldsymbol{x}) = \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{\boldsymbol{\epsilon}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}\left[\|\boldsymbol{\epsilon}-\boldsymbol{\epsilon}_\theta(\mathbf{x_t},t)\|^2\right]$$

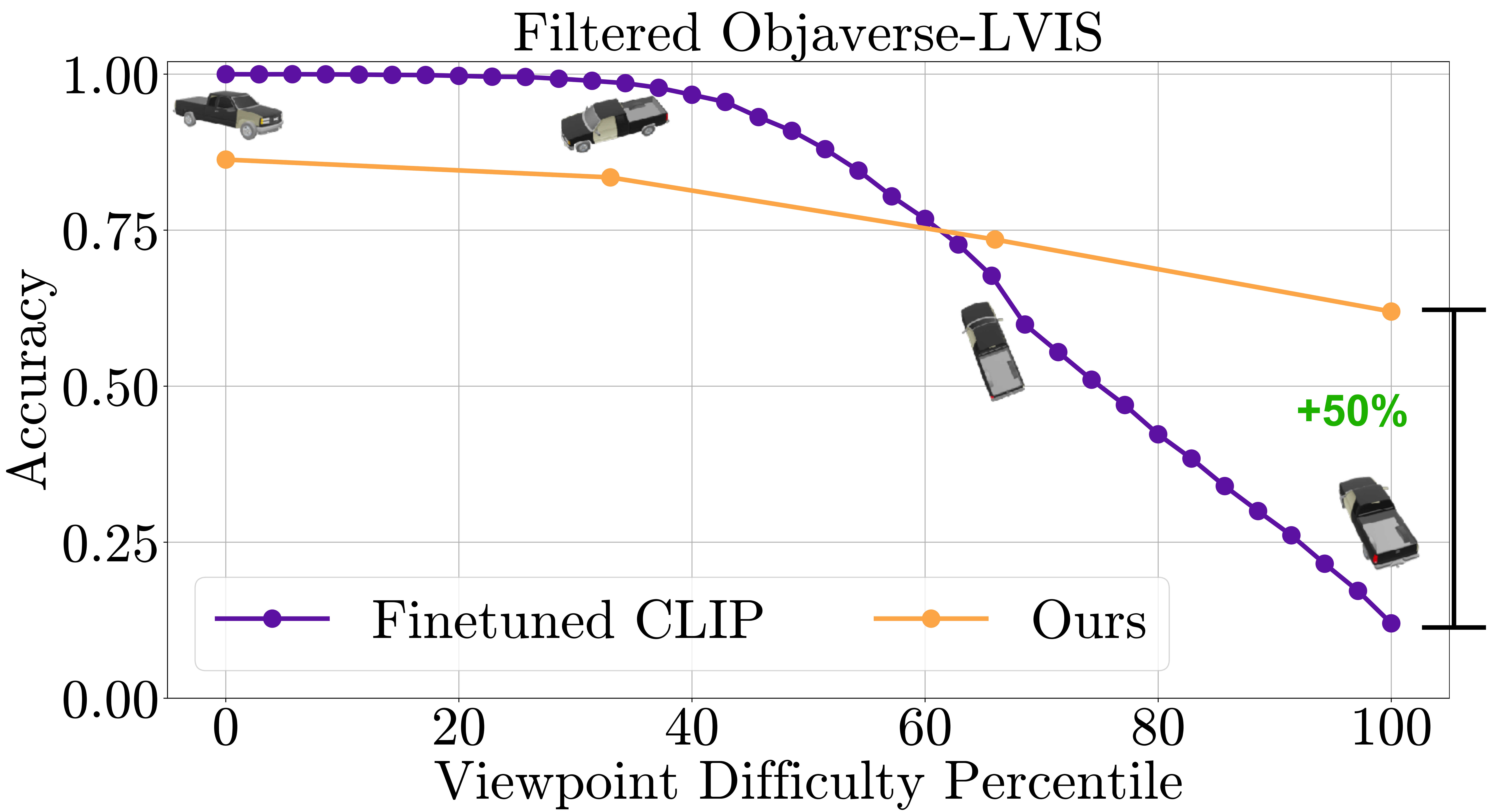
Diffusion Energy

"Distribution Slices": Why FoCal Generalizes Across Transforms



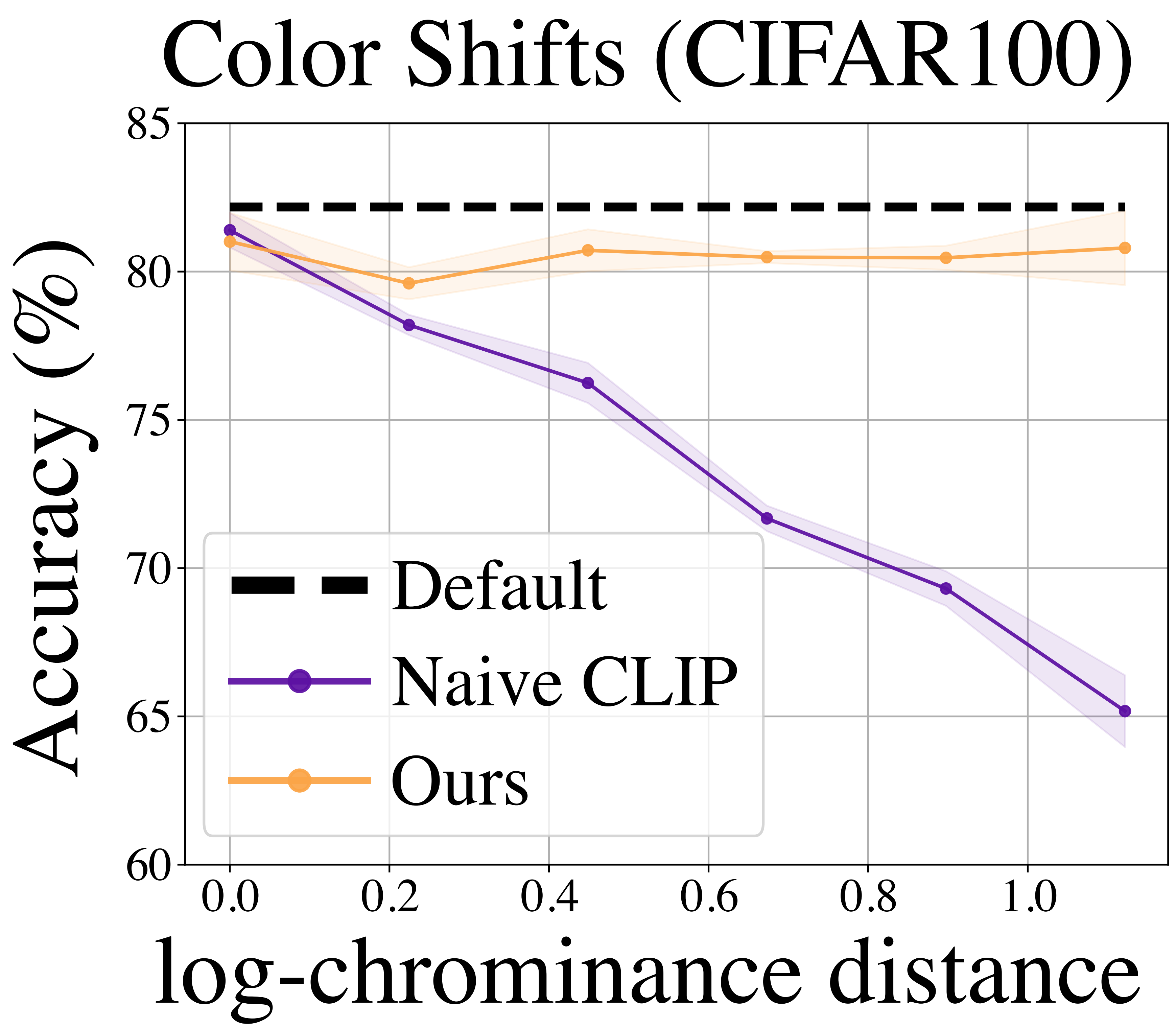
Insight: Transformed images form a "slice" of the natural image distribution, and foundation models have already learned a prior over this distribution.

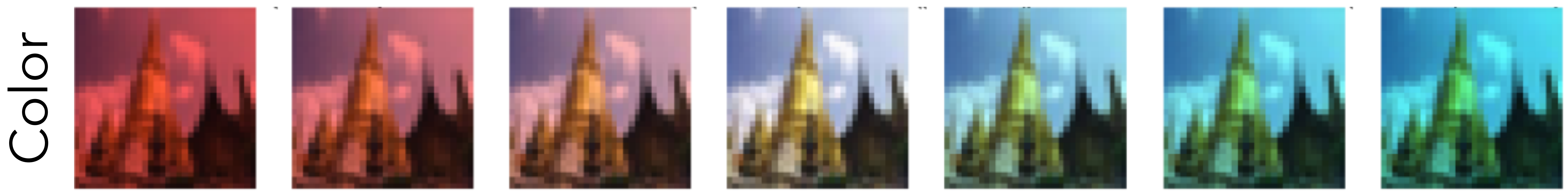
Significant Improvement on Worst 3D Viewpoints



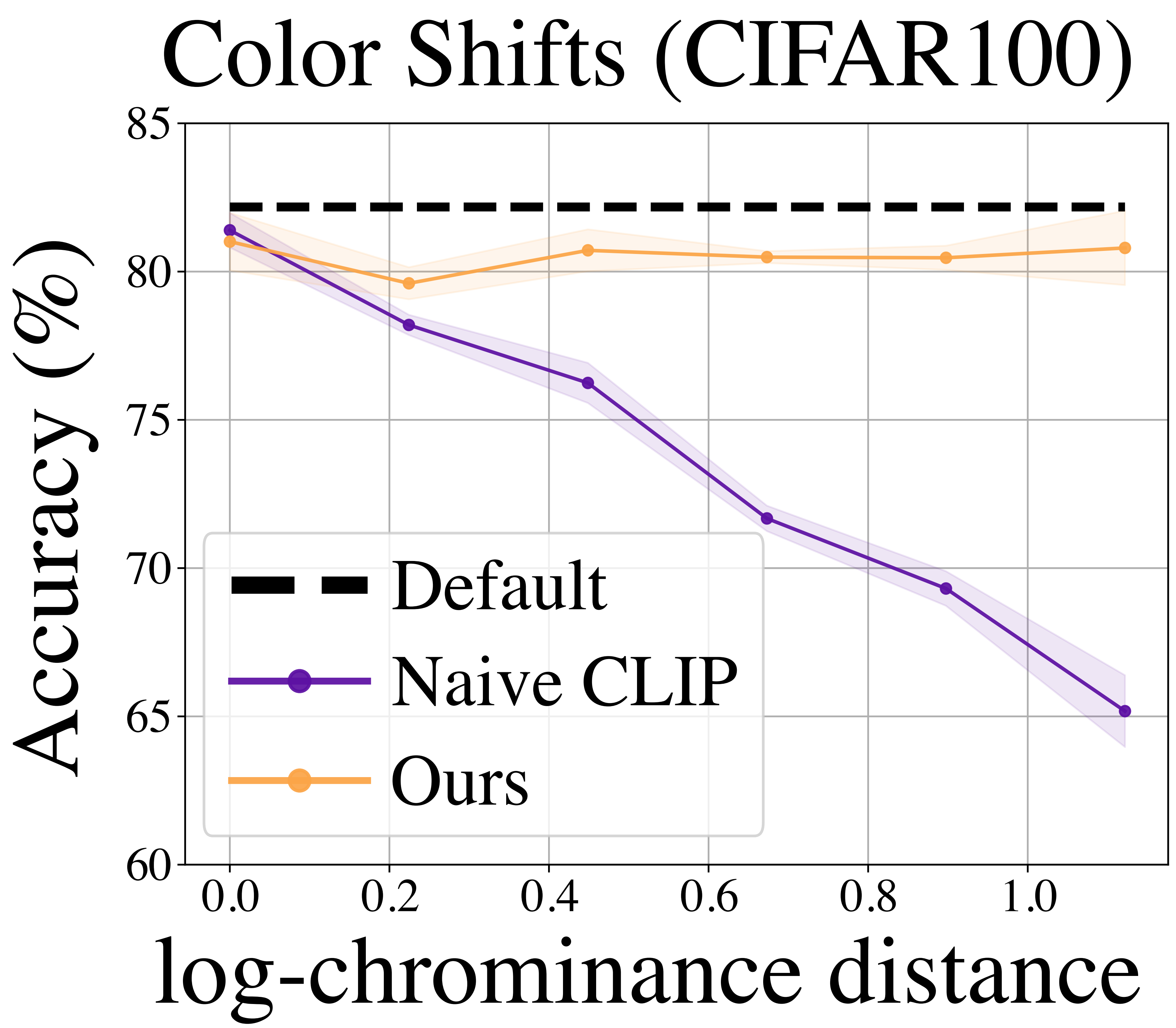
Significant Robustness Boost for Color and Contrast

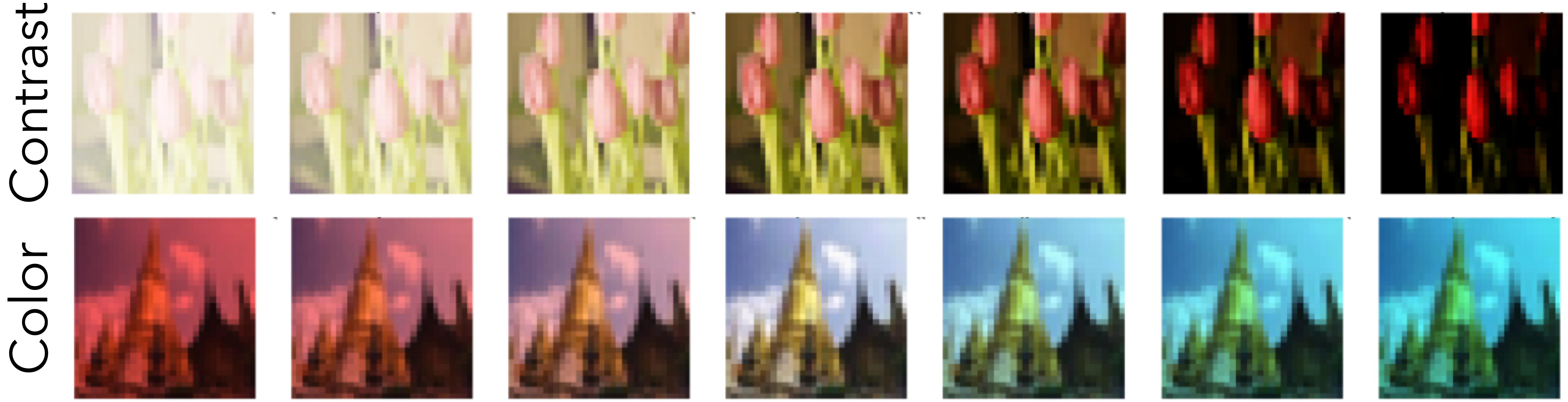
Significant Robustness Boost for Color and Contrast



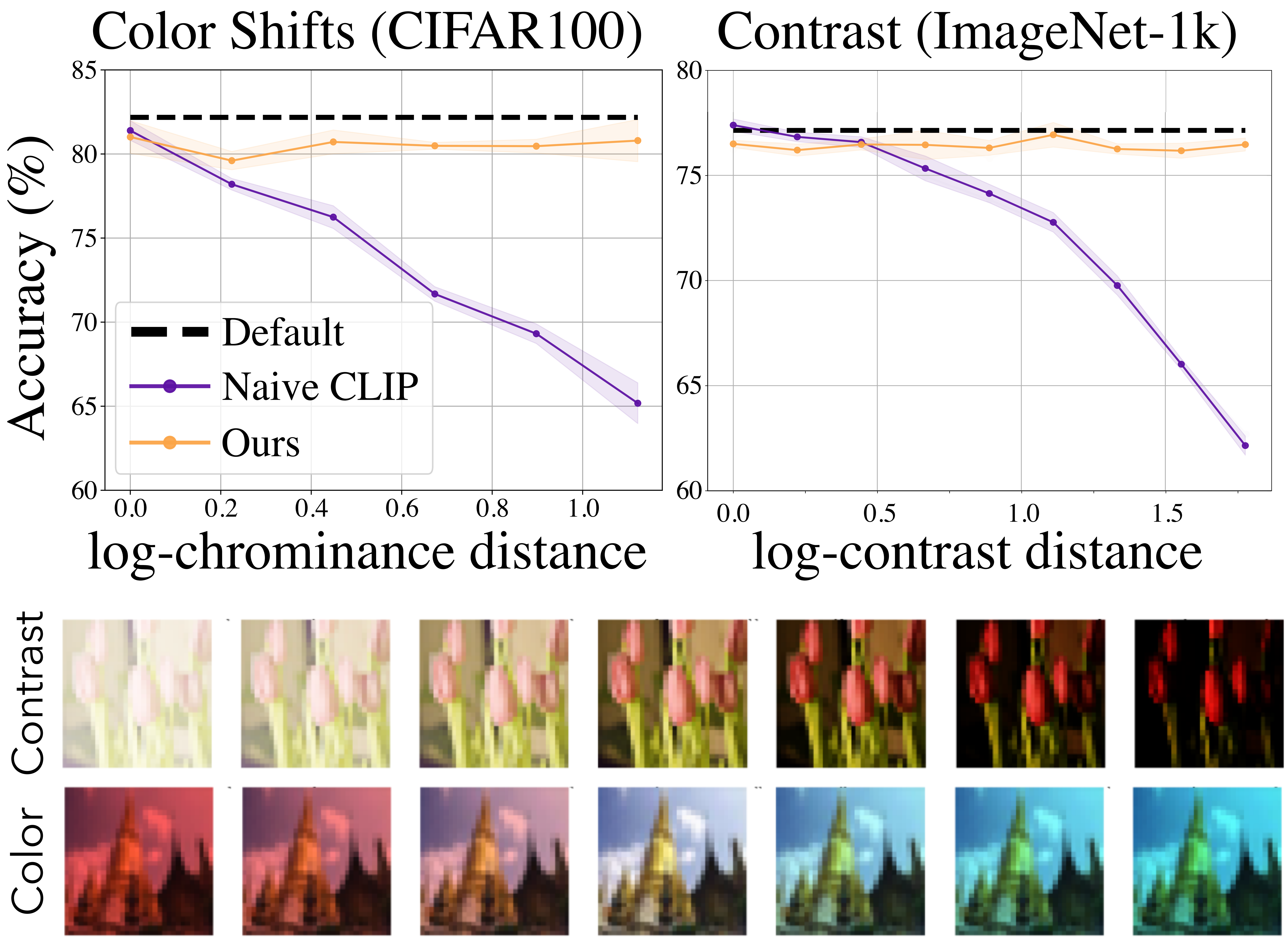


Significant Robustness Boost for Color and Contrast





Significant Robustness Boost for Color and Contrast



Day-Night Results

Night Image

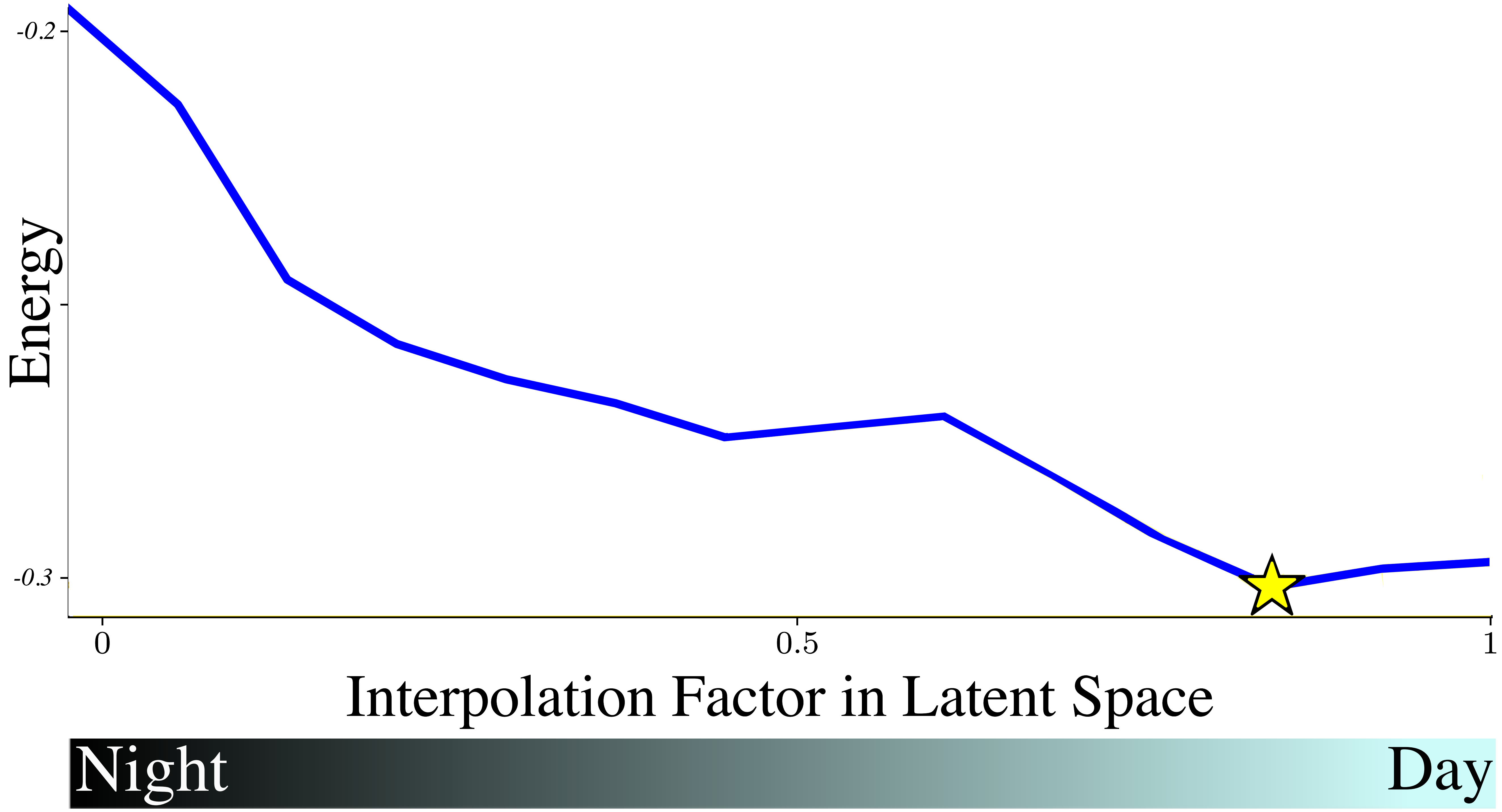
Day Image

Day-Night Results

Night Image

Day Image

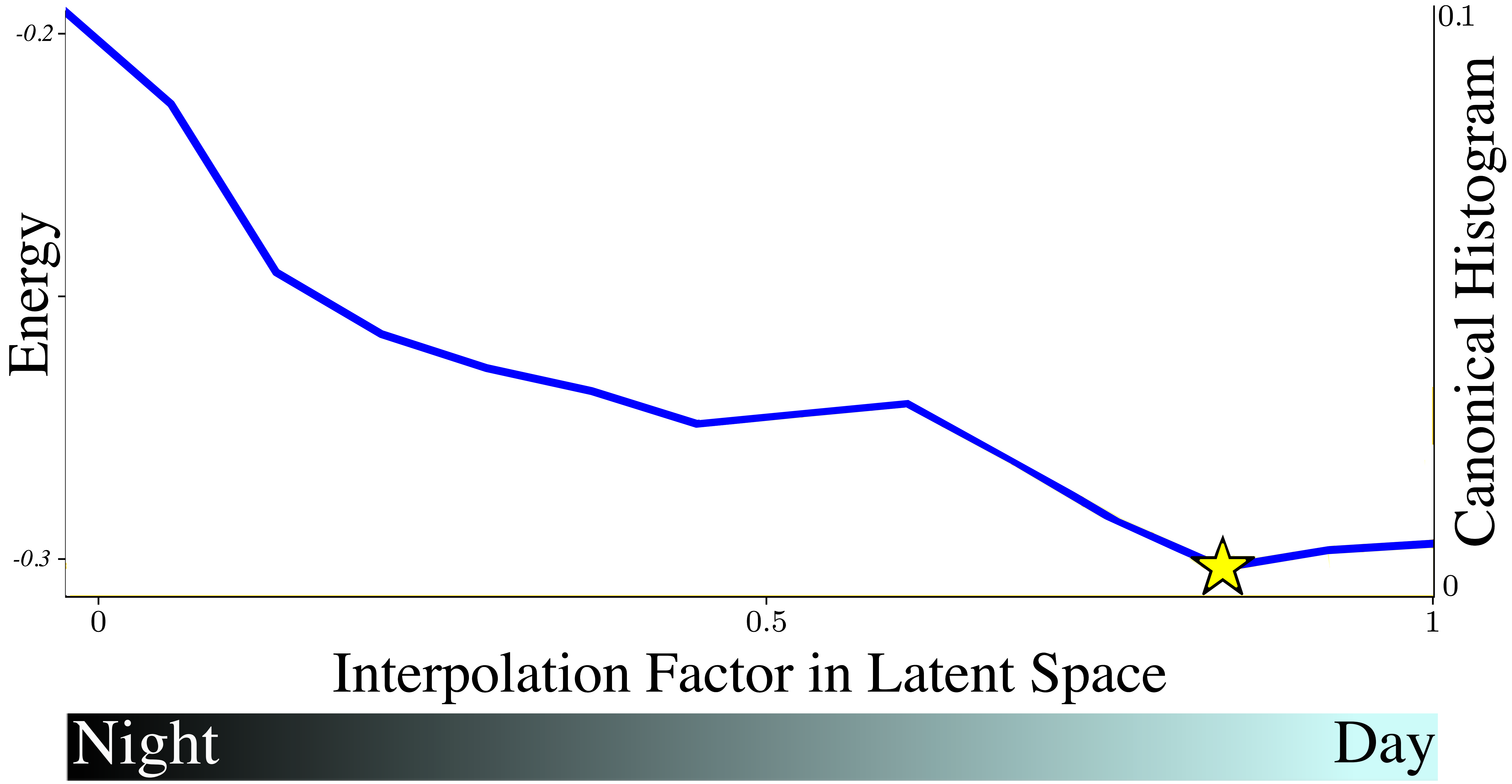
Day-Night Results



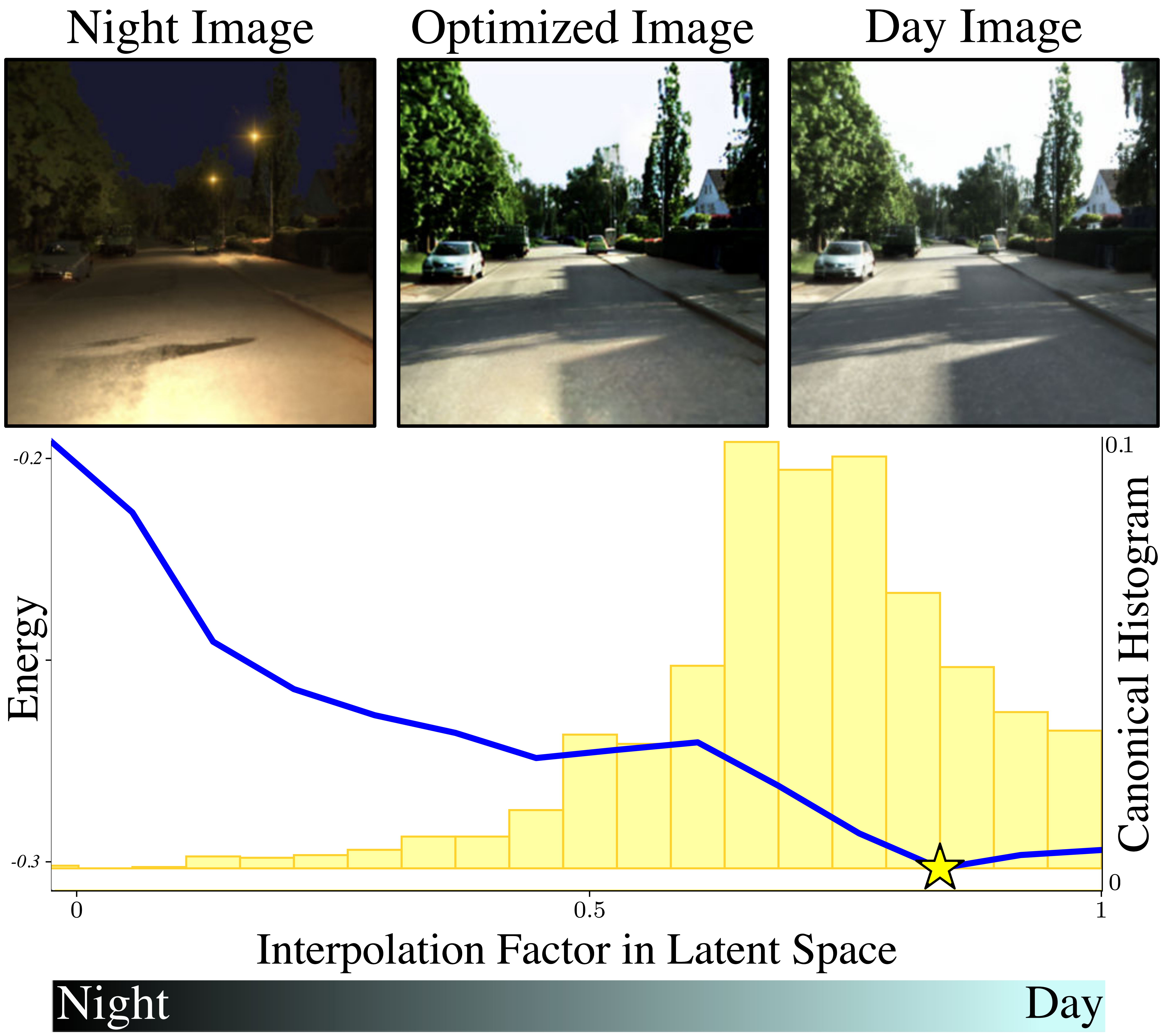
Day-Night Results

Night Image

Day Image

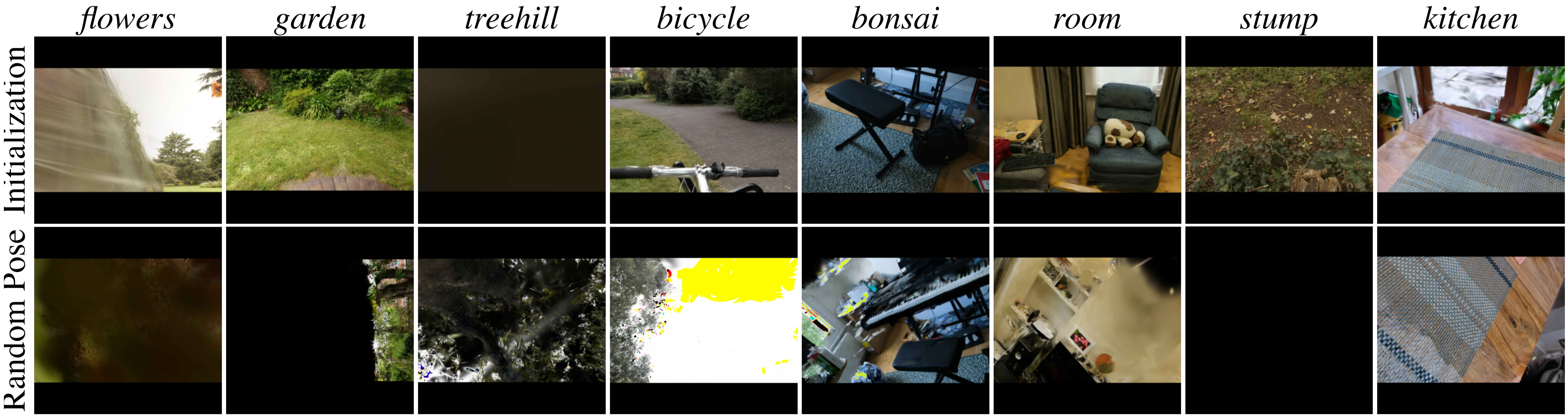


Day-Night Results



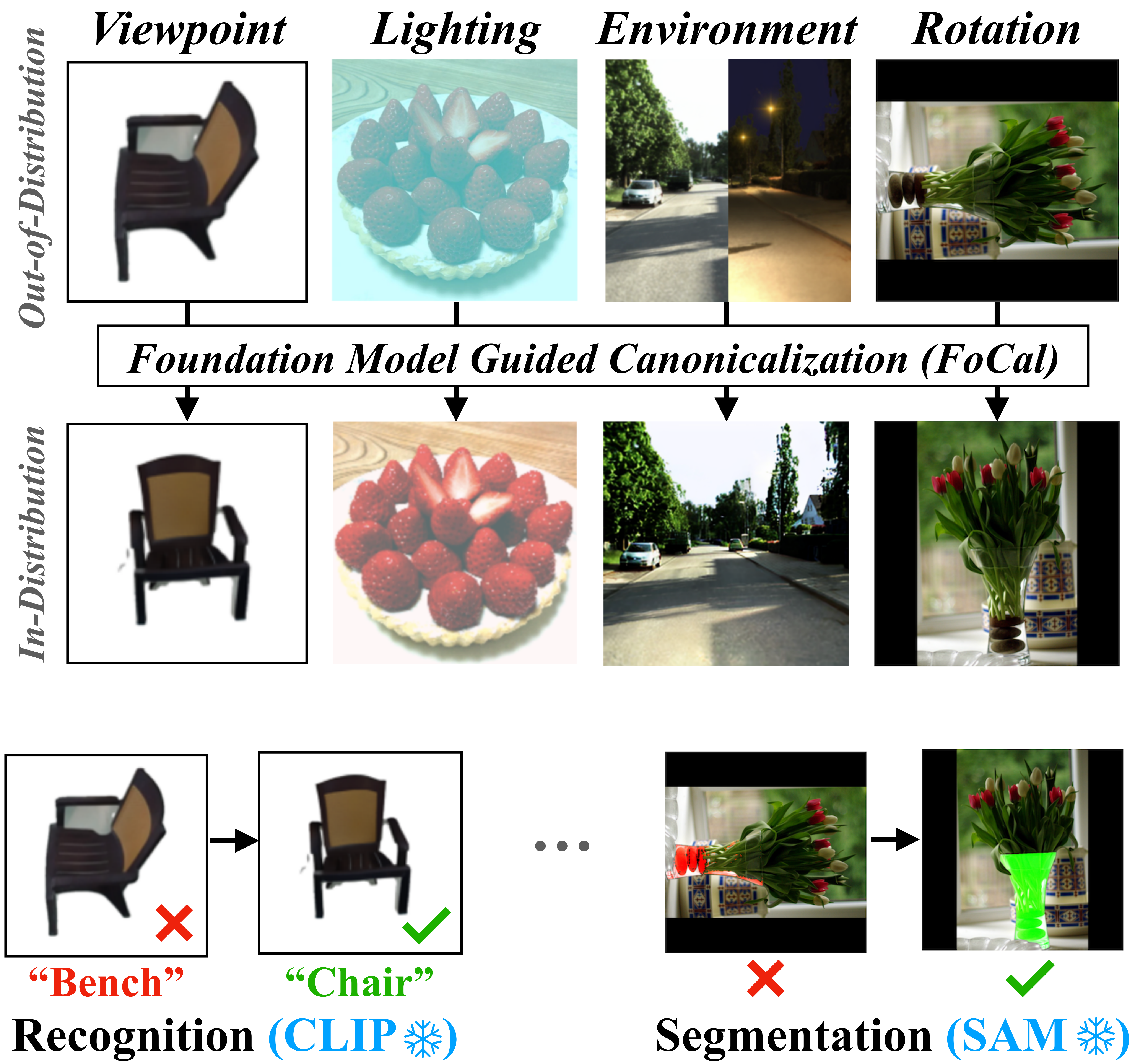
Active Vision Results

Active Vision Results



Active Vision Results

FoCal: a *Scalable, Data-Driven, Test-Time* Approach to Robust Perception



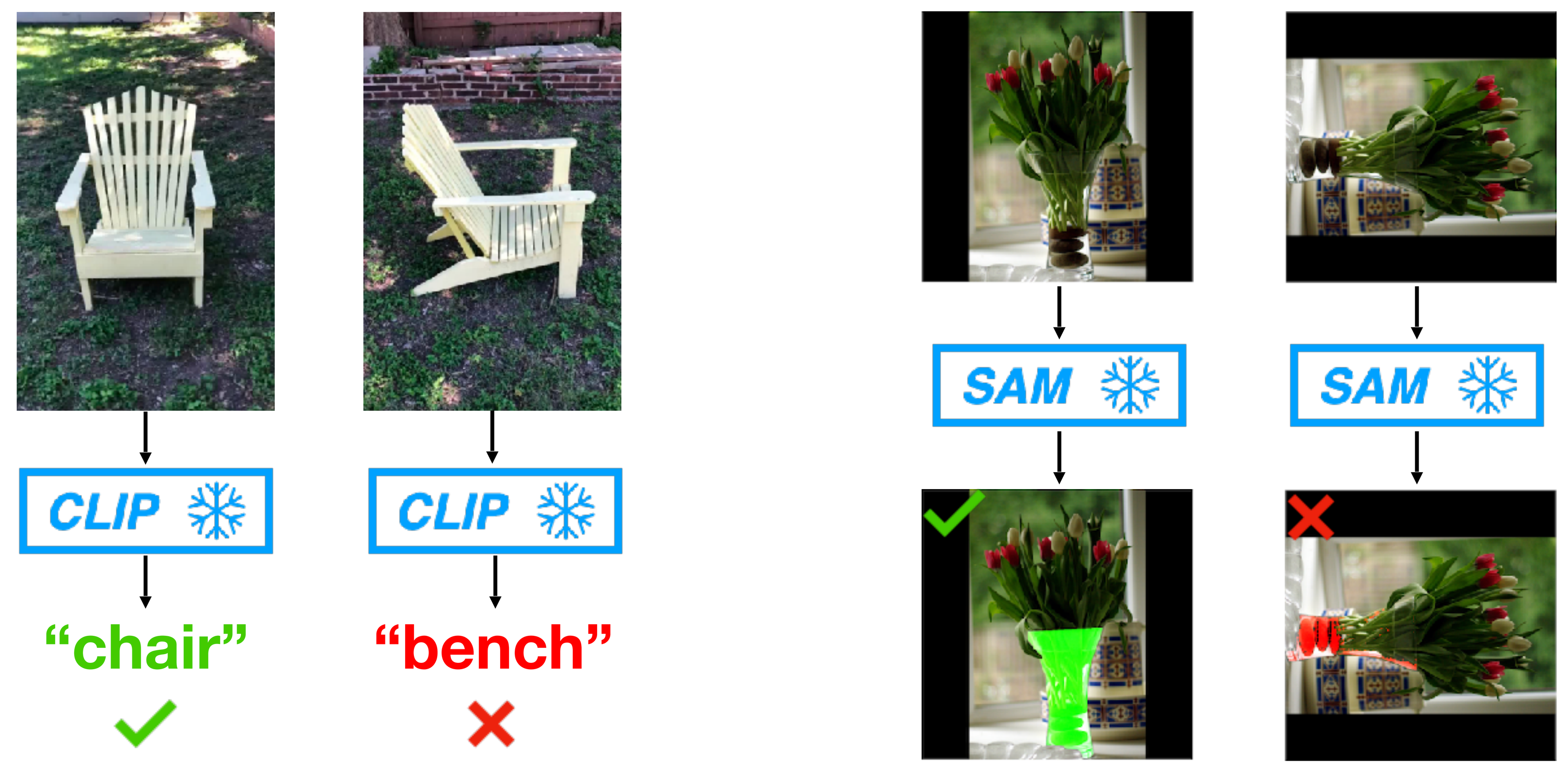
Test-Time Canonicalization by Foundation Models for Robust Perception

Utkarsh Singhal* Ryan Feng* Stella X. Yu Atul Prakash

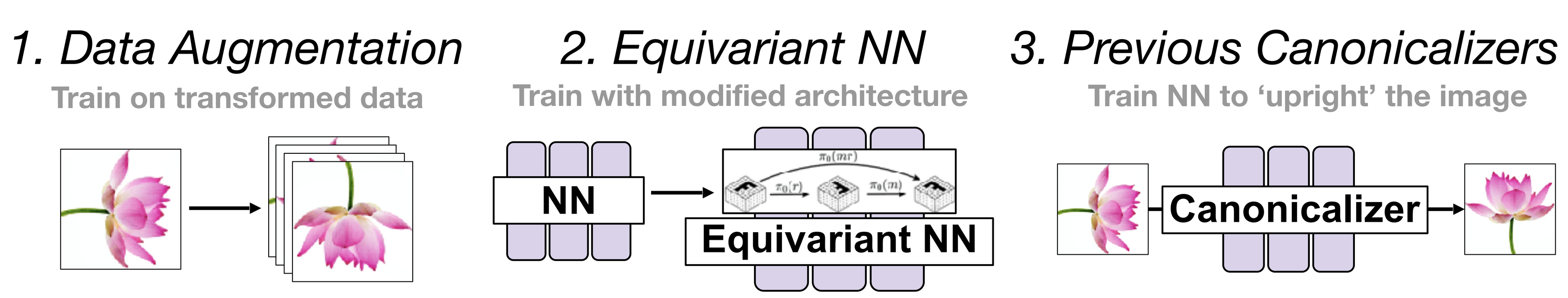
Summary: *Test-time search makes models more robust to natural input variations by converting the varied versions of the input into a 'typical' version.*

Motivation

Foundation models (FM) are *brittle*



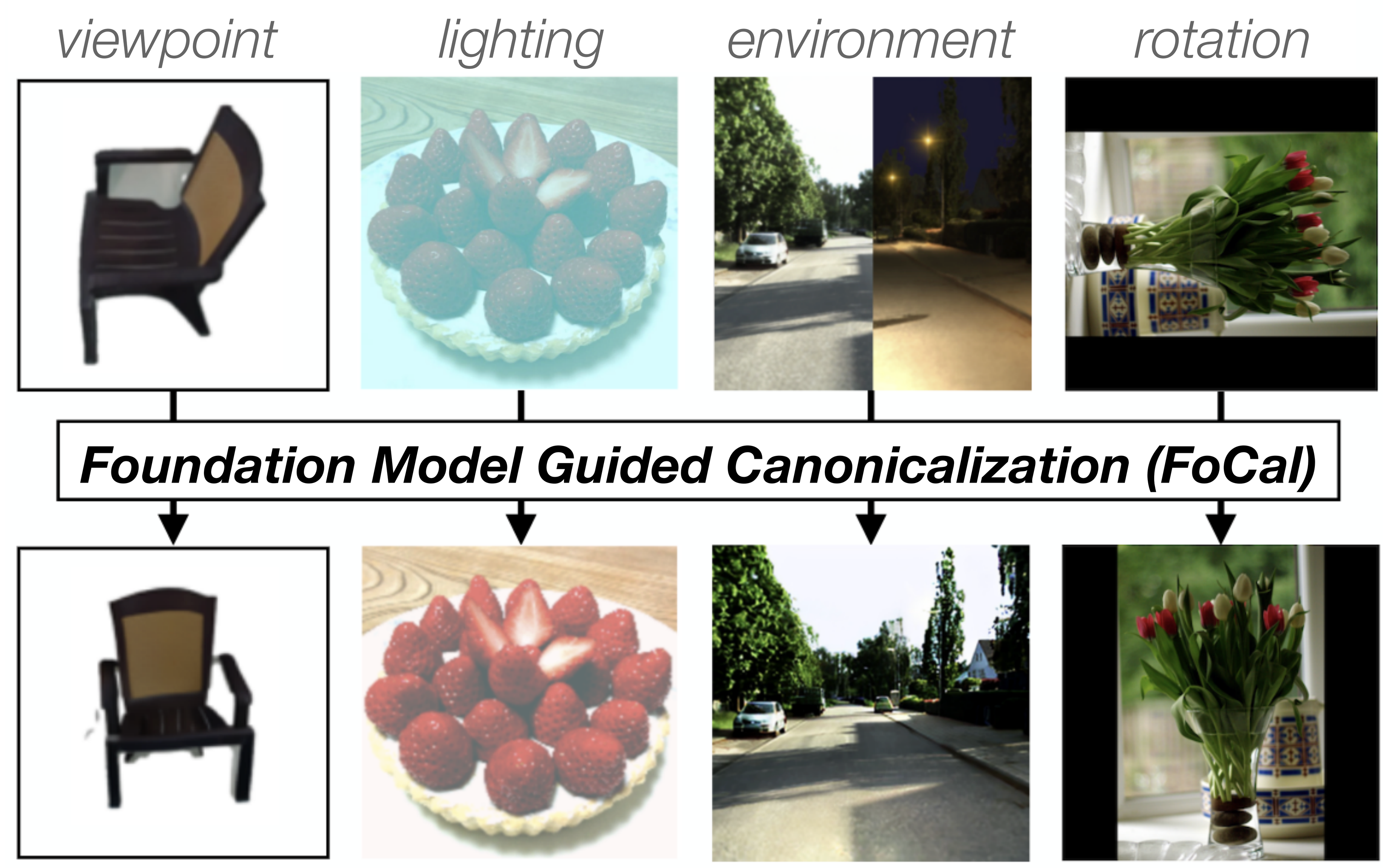
Prior approaches rely on *transform-specific training*



***Problem**: These methods struggle with OOD input variations at test-time*

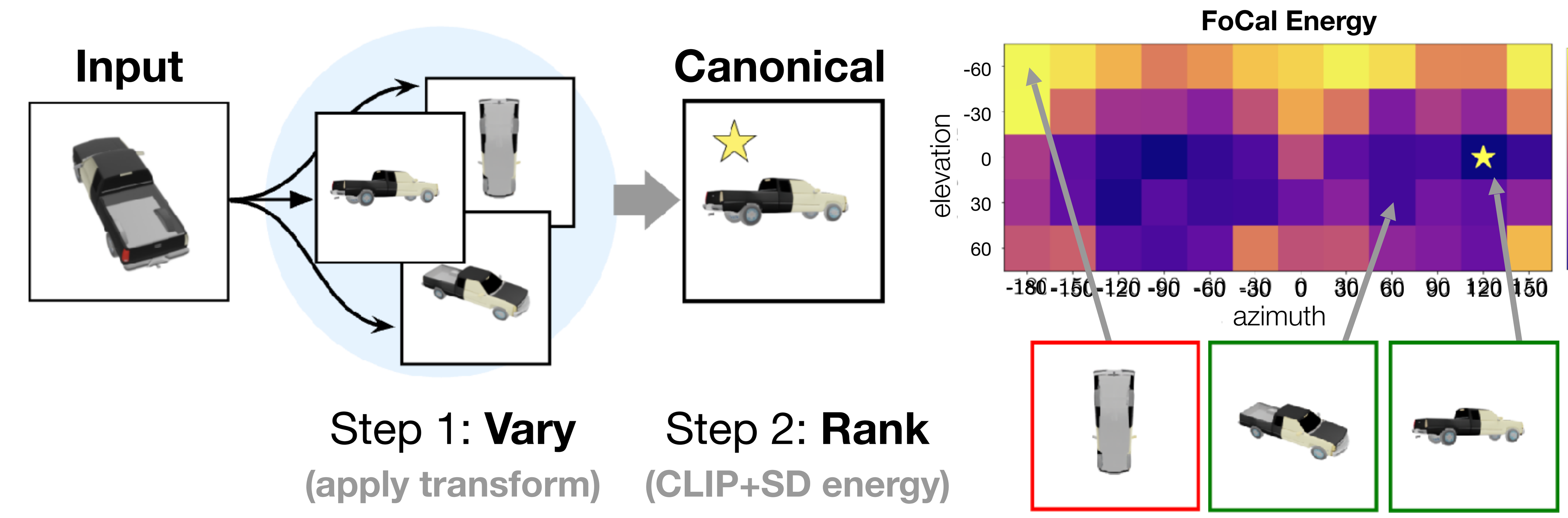
FoCal: Test-Time Canonicalization

Idea: transform input to the most 'typical' version

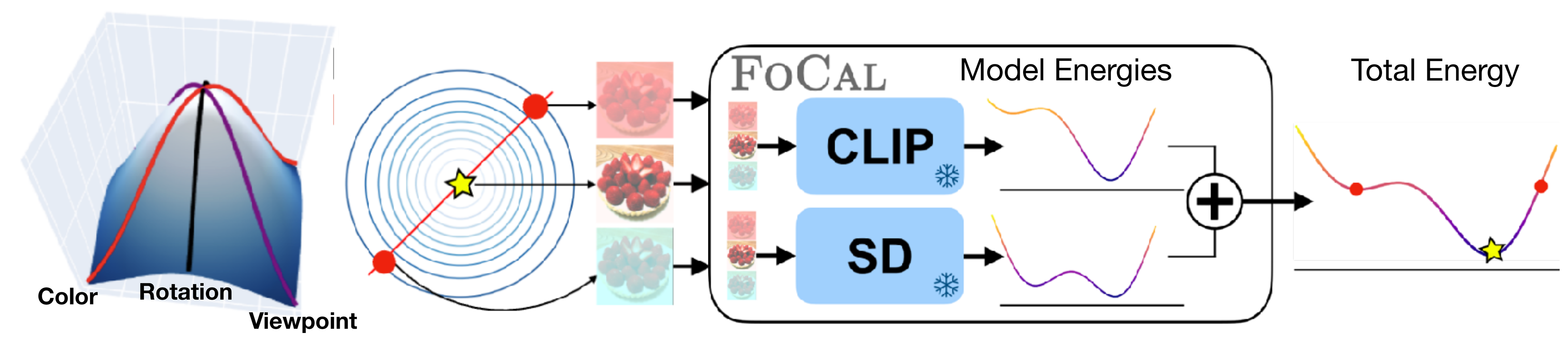


+ *Robustness for many natural variations*
+ *Any downstream task, any model, no training*
+ *Guaranteed invariance for invertible transforms*

Approach: Vary & Rank



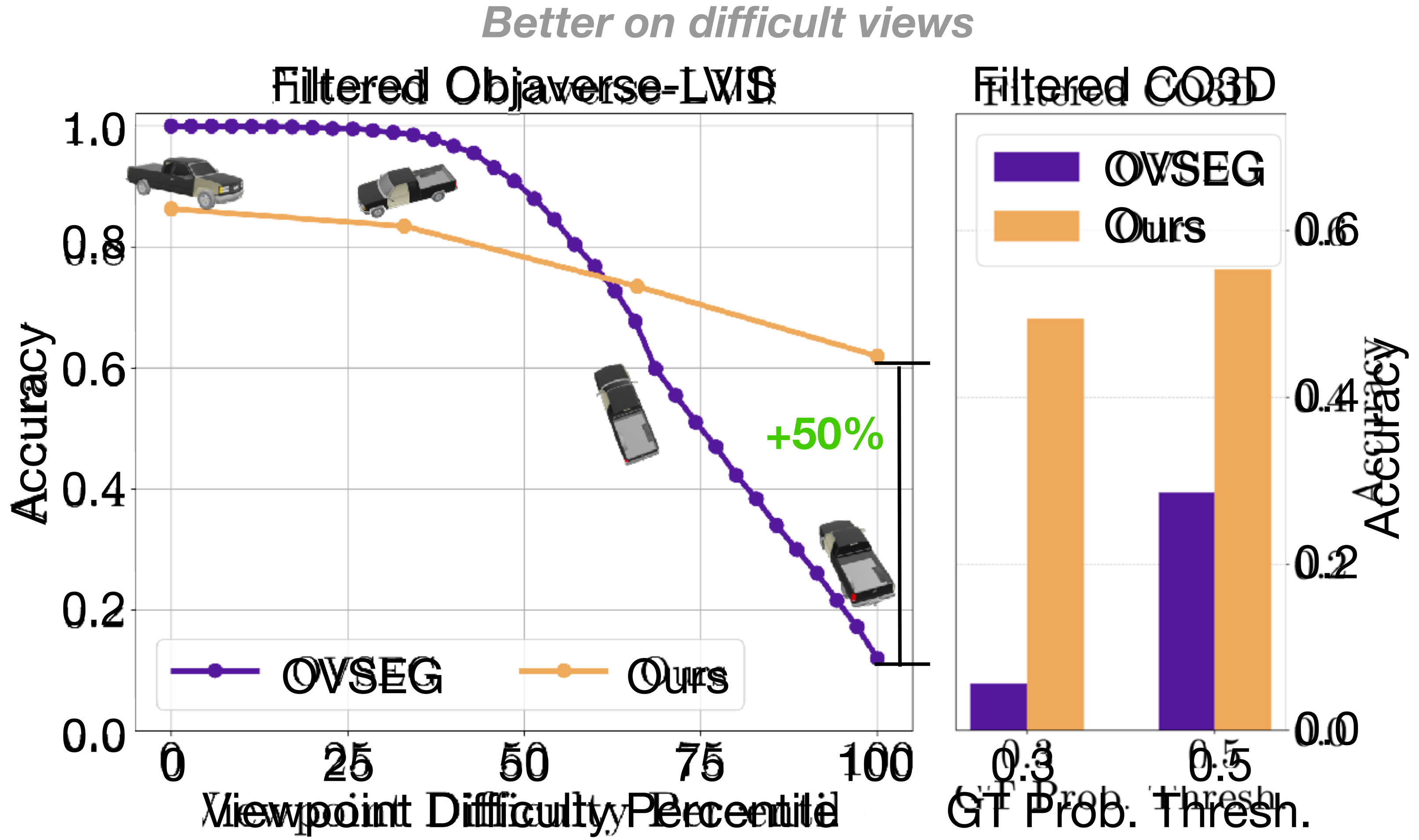
Insight: FM energy estimates the input 'typicality' for many natural variations. Minimizing FM energy over a transform yields robustness.



Results: FoCal Improves Robustness Across Many Domains

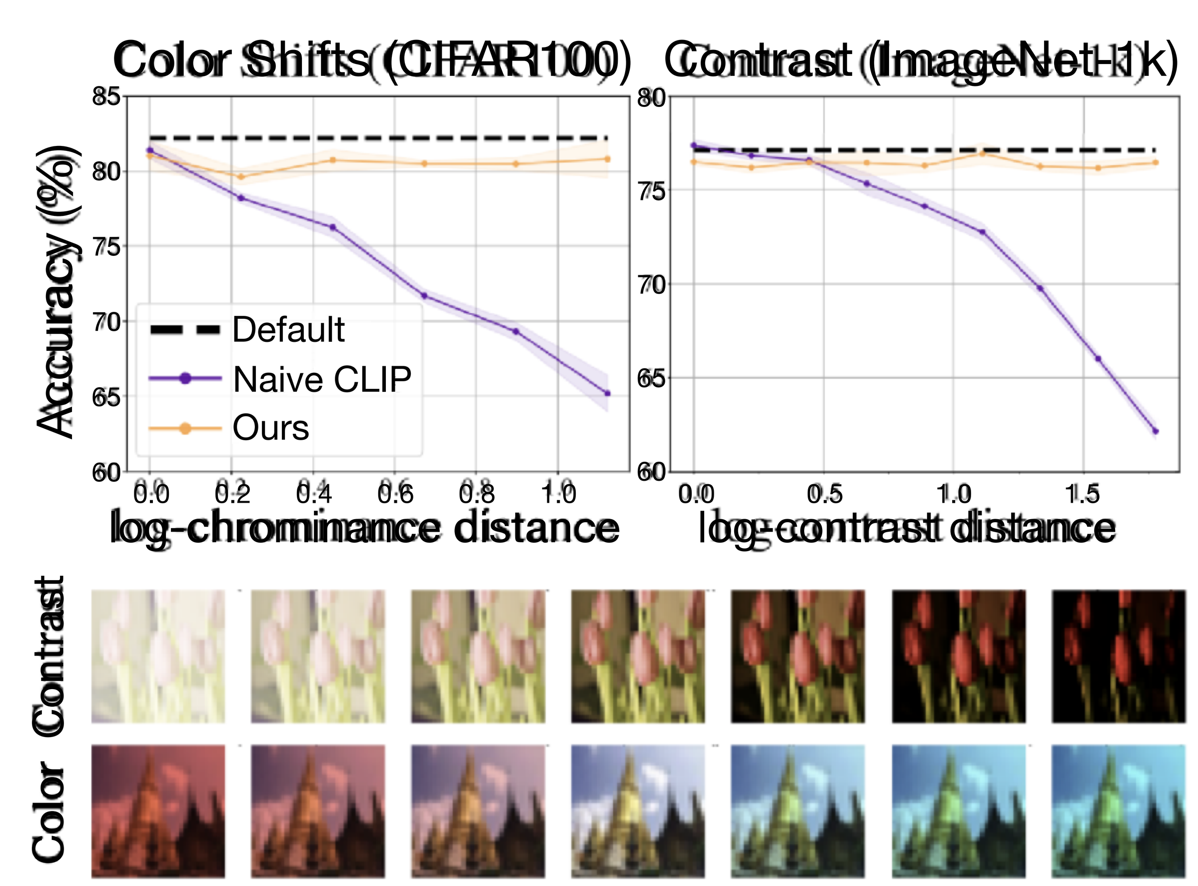
3D Viewpoint Shifts

Better on difficult views



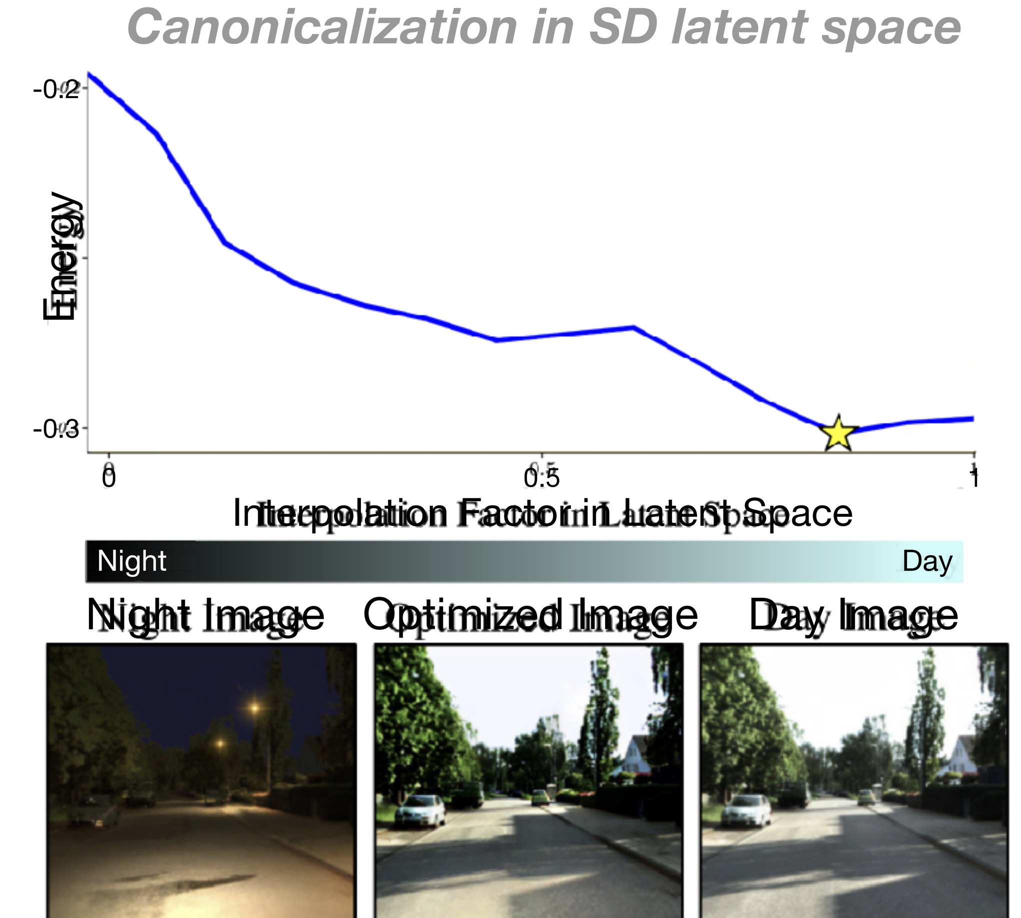
Color and Contrast

Better robustness across color / contrast shifts



Day-Night

Canonicalization in SD latent space



Active Vision (Exploring Virtual Environment)

FoCal looks at salient objects in upright poses