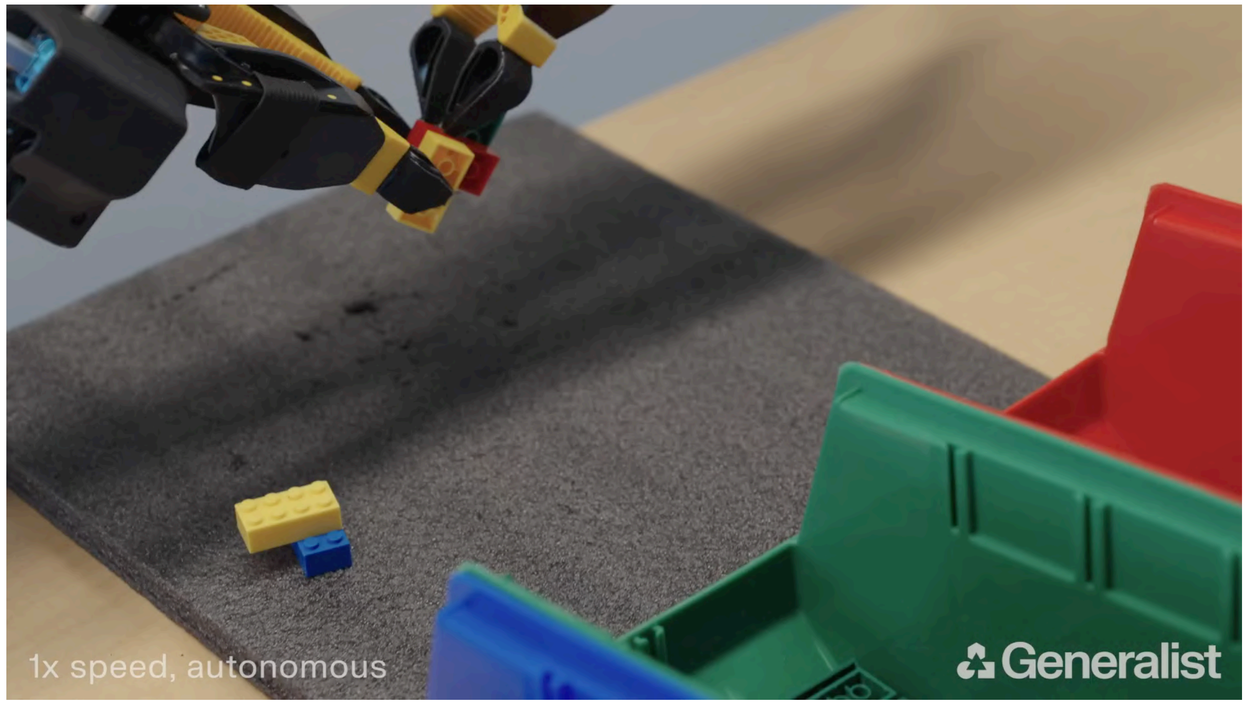
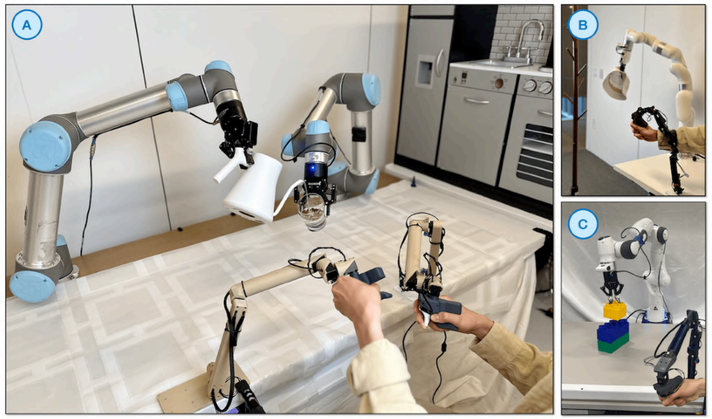
CSCI 699: Robotic Perception

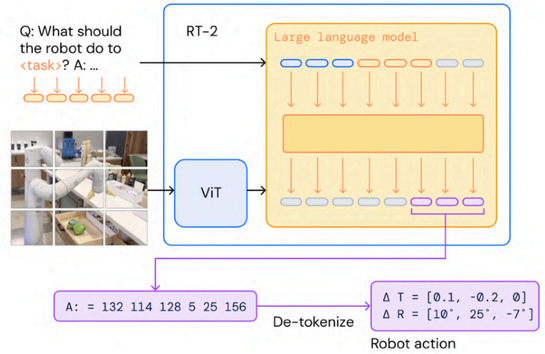
Yue Wang

Aug 27th, 2025



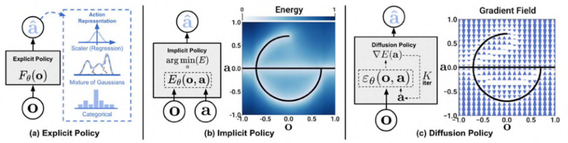
Cambrian Explosion of Robotics





Diffusion Policy

Visuomotor Policy Learning via Action Diffusion



This course

- Computer vision + robotics
- Focus on visual-motor policy learning
- Practice driven, a semester long project, fast paced, weekly project update
- Submit a paper to RSS/ICML/IROS
- Suitable for who want to do robotic research, not suitable for who want to take a regular computer vision / robotic course
- We will occasionally have invited speakers who are experts on robotics

Plan for the course (subject to changes)

- Weeks 1-3 Foundations of Computer Vision and Robotics
- Weeks 4 -13 Project
- Week 4 Project Proposal Due
- Week 8 Midterm Report Due
- Week 13 Final Presentation
- Week 14 Final Report Due

Grading

- Class participation (10%). Attendance is required. You may miss up to two sessions without penalty. Each additional unexcused absence will result in a deduction of 1% from the final grade
- Project proposal (5%)
- Weekly project update (20%)
- Midterm report (15%)
- Final presentation (20%)
- Final report (30%)
- Five grace days in total that may be applied to the midterm report or the final report. No more than three grace days can be used on each.

Project

- Simulation
- Perception
- Manipulation with robot arms
- Locomotion with quadrupeds
- Humanoid robots
- Must have vision components (state-based only is not allowed), must be deployable to real robots.

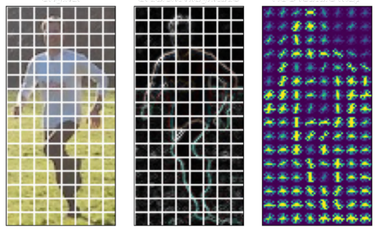
Project

- Group of 4-6 students.
- Work must be equally distributed.
- Each group gives a 25-minute update, starting week 5. Each student has to present at least once.
- We can provide physical robots.

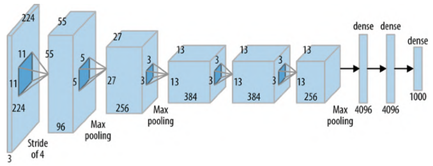
Robots

Data is the key to artificial intelligence.

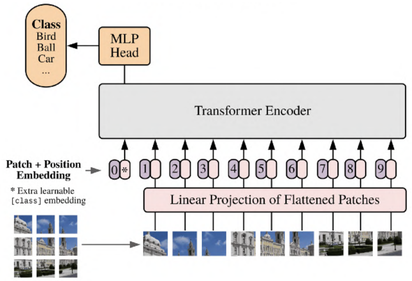
Physical AI



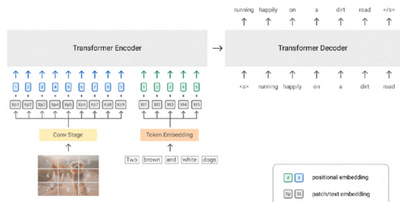
HOG+SIFT+SVM



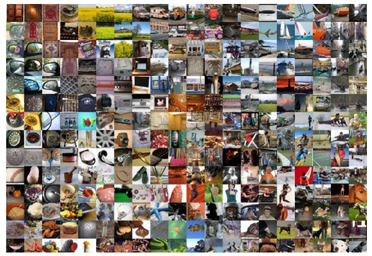
CNNs

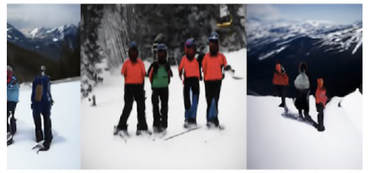


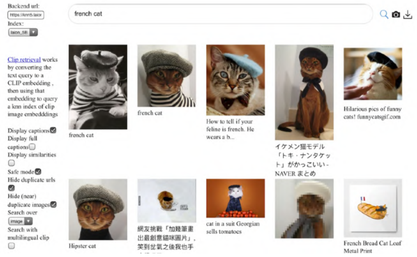
Transformers

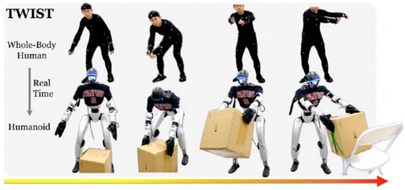


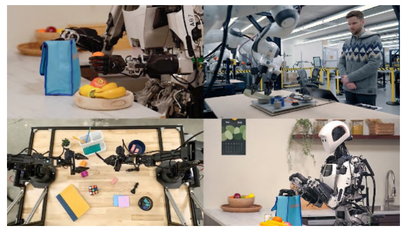
VLMs

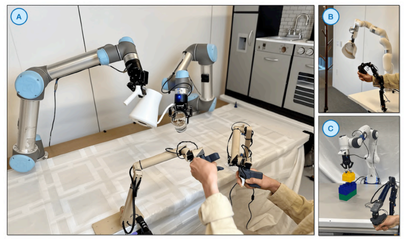




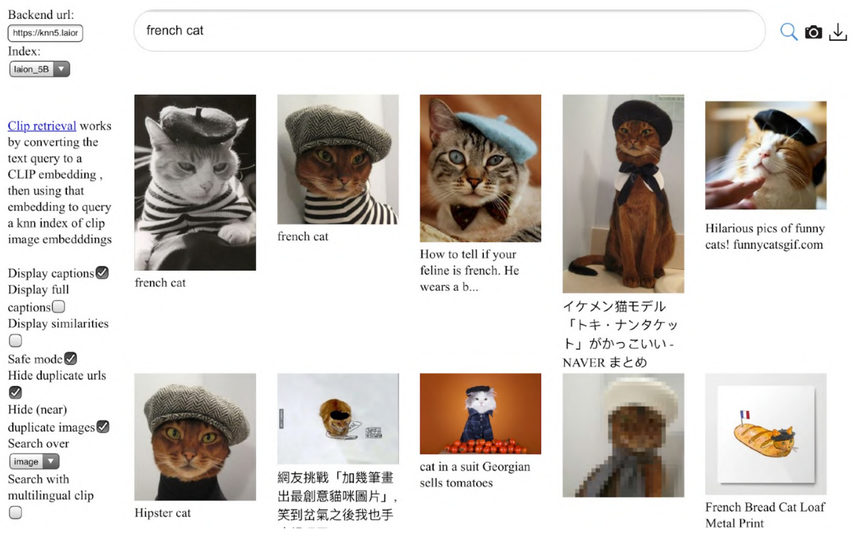


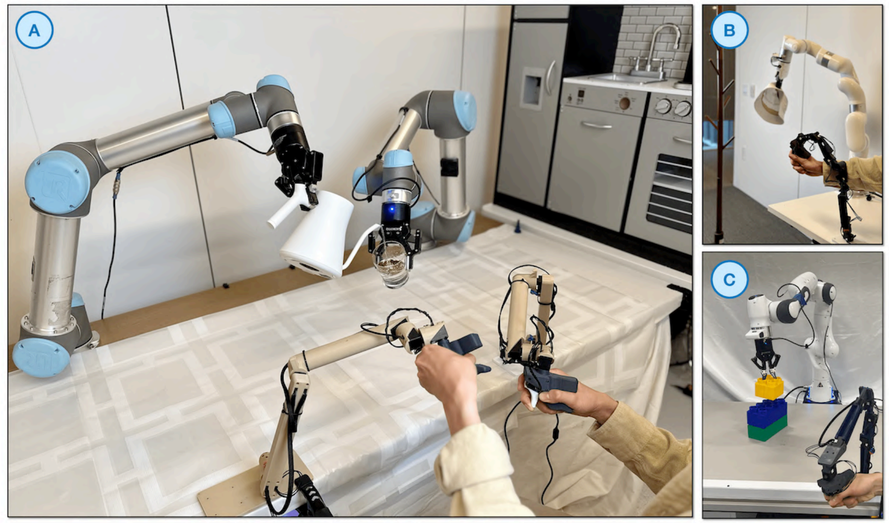






Little Data | Curated | Web Scale | Noisy Multimodal | Embodied Data





< 1s

Ubiquitous

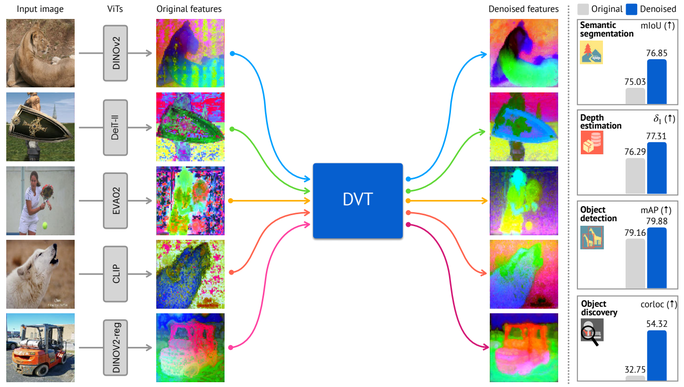
$0.01 per data point

> 60s

Confined to lab environments

$5 per data point

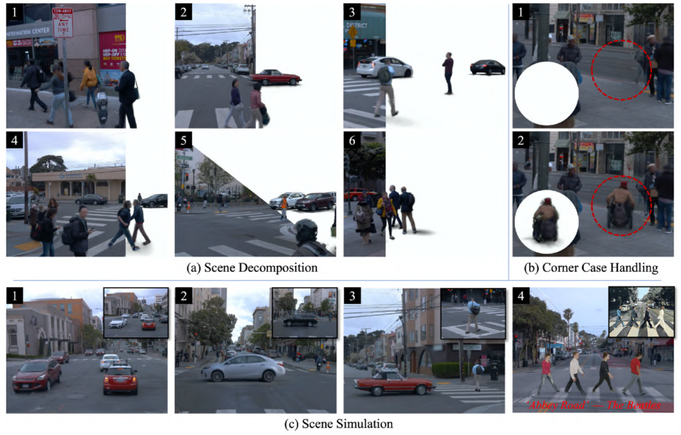
How to scale up embodied data with minimal human supervision?

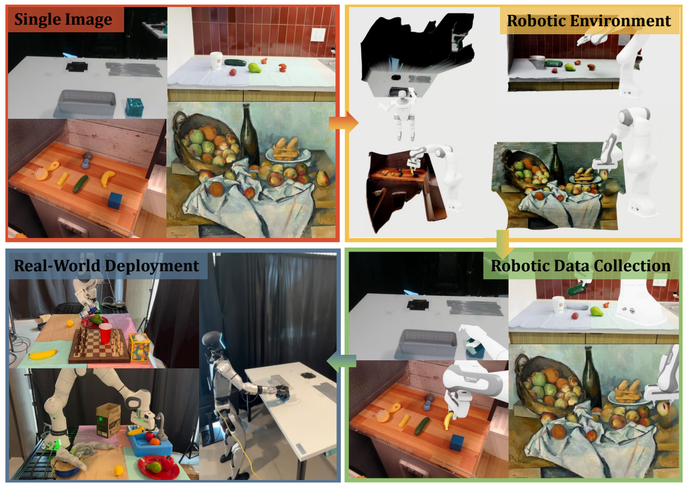


[ECCV 2024] "Denoising Vision Transformers." Yang et al.

[ICLR 2025] "PhysBench: Benchmarking and Enhancing Vision-Language Models for Physical World Understanding." Zhou et al.

Foundation models for vision and physics.

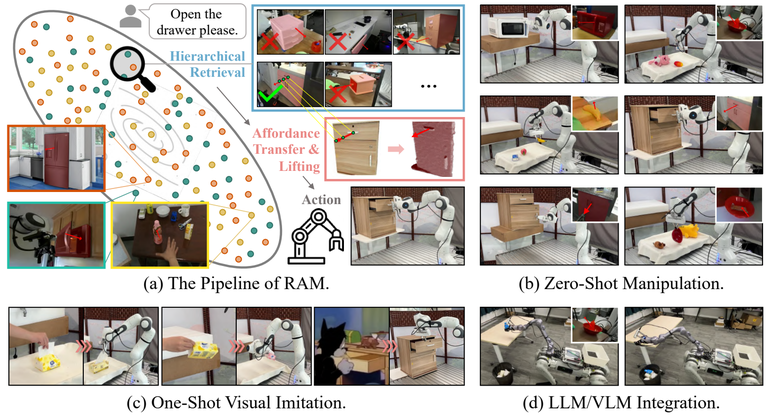




[ICLR 2025] "Omni Urban Scene Reconstruction." Chen et al.

[In submission 2025] "Robot Learning from Any Images" Zhao et al.

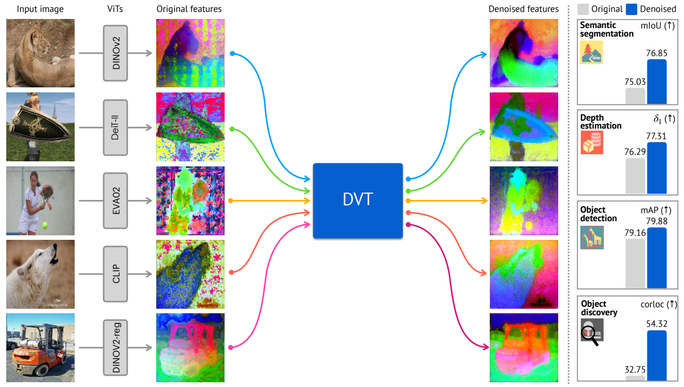
3D reconstruction for physical simulation and generation.



[CoRL 2024] "RAM: Retrieval-Based Affordance Transfer for Generalizable Zero-Shot Robotic Manipulation." Kuang et al.

[In submission 2025] "Learning from Massive Human Videos for Universal Humanoid Pose Control." Mao et al.

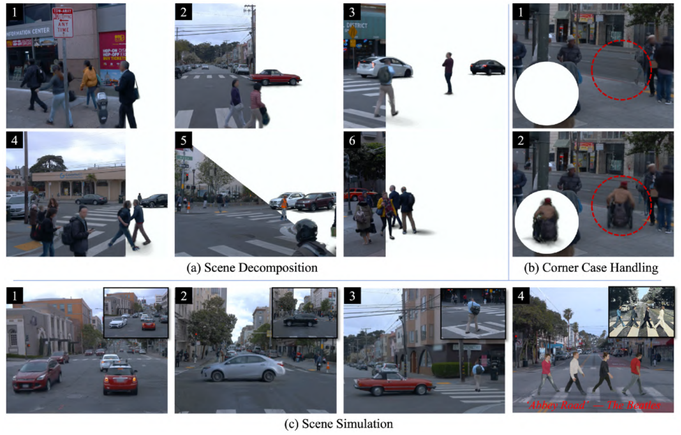
Robot learning from non-robot data.

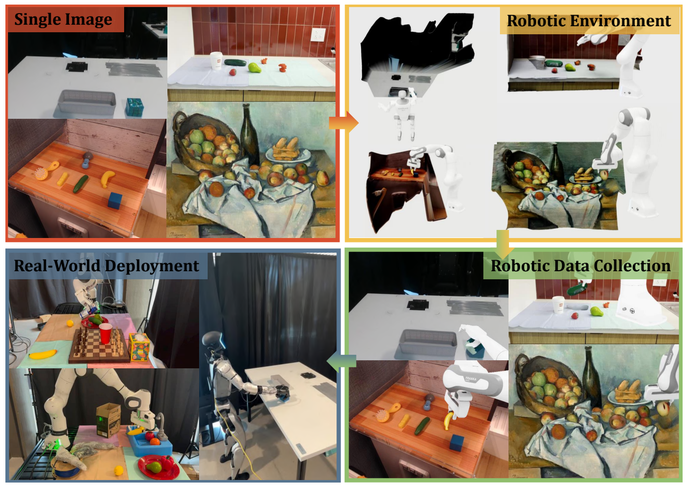


[ECCV 2024] "Denoising Vision Transformers." Yang et al.

[ICLR 2025] "PhysBench: Benchmarking and Enhancing Vision-Language Models for Physical World Understanding." Zhou et al.

Foundation models for vision and physics.

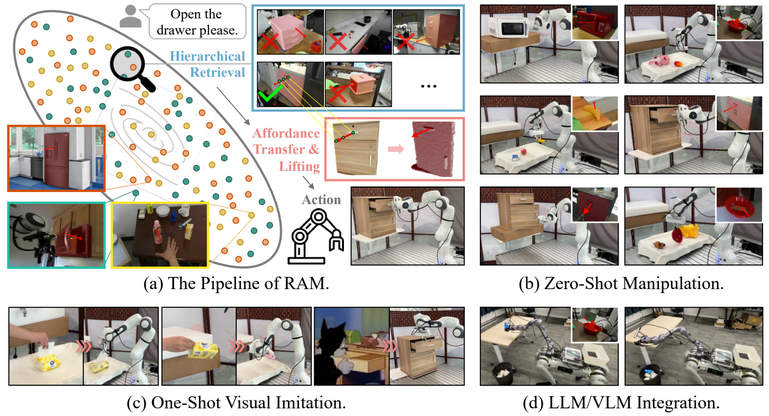




[ICLR 2025] "Omni Urban Scene Reconstruction." Chen et al.

[In submission 2025] "Robot Learning from Any Images" Zhao et al.

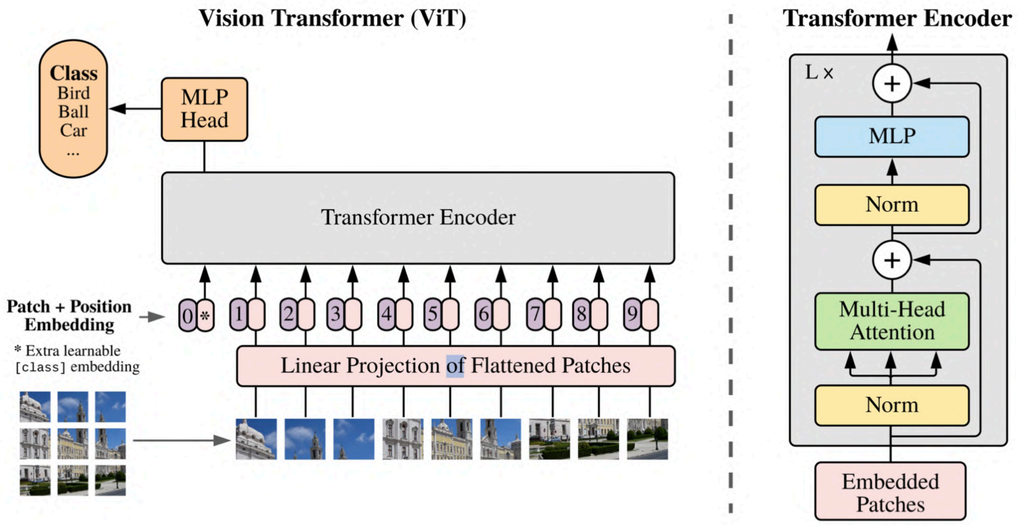
3D reconstruction for physical simulation and generation.



[CoRL 2024] "RAM: Retrieval-Based Affordance Transfer for Generalizable Zero-Shot Robotic Manipulation." Kuang et al.

[In submission 2025] "Learning from Massive Human Videos for Universal Humanoid Pose Control." Mao et al.

Robot learning from non-robot data.



Are ViTs good enough?

Vision Transformers (ViTs)

Classification

Object Detection

Semantic Segmentation

Depth Estimation

Point Tracking

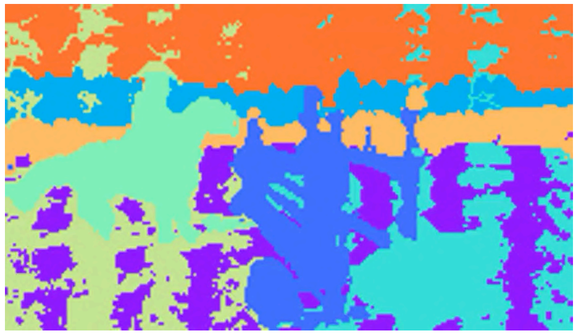
Object Discovery

Artifacts in ViTs: DINOv2 as an example

Input

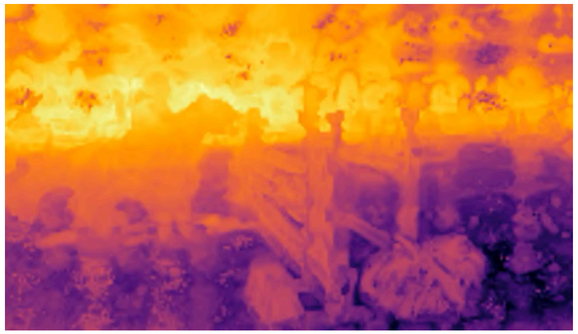
DINOv2

Feature Map PCA

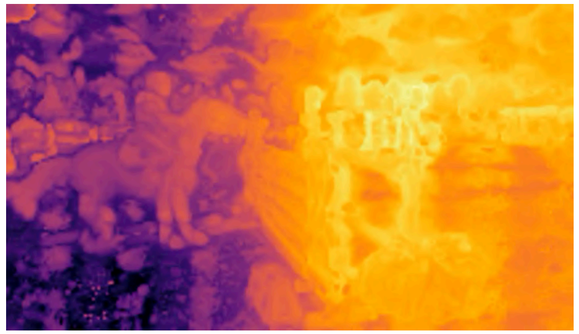


Noisy Clusters

PCA1



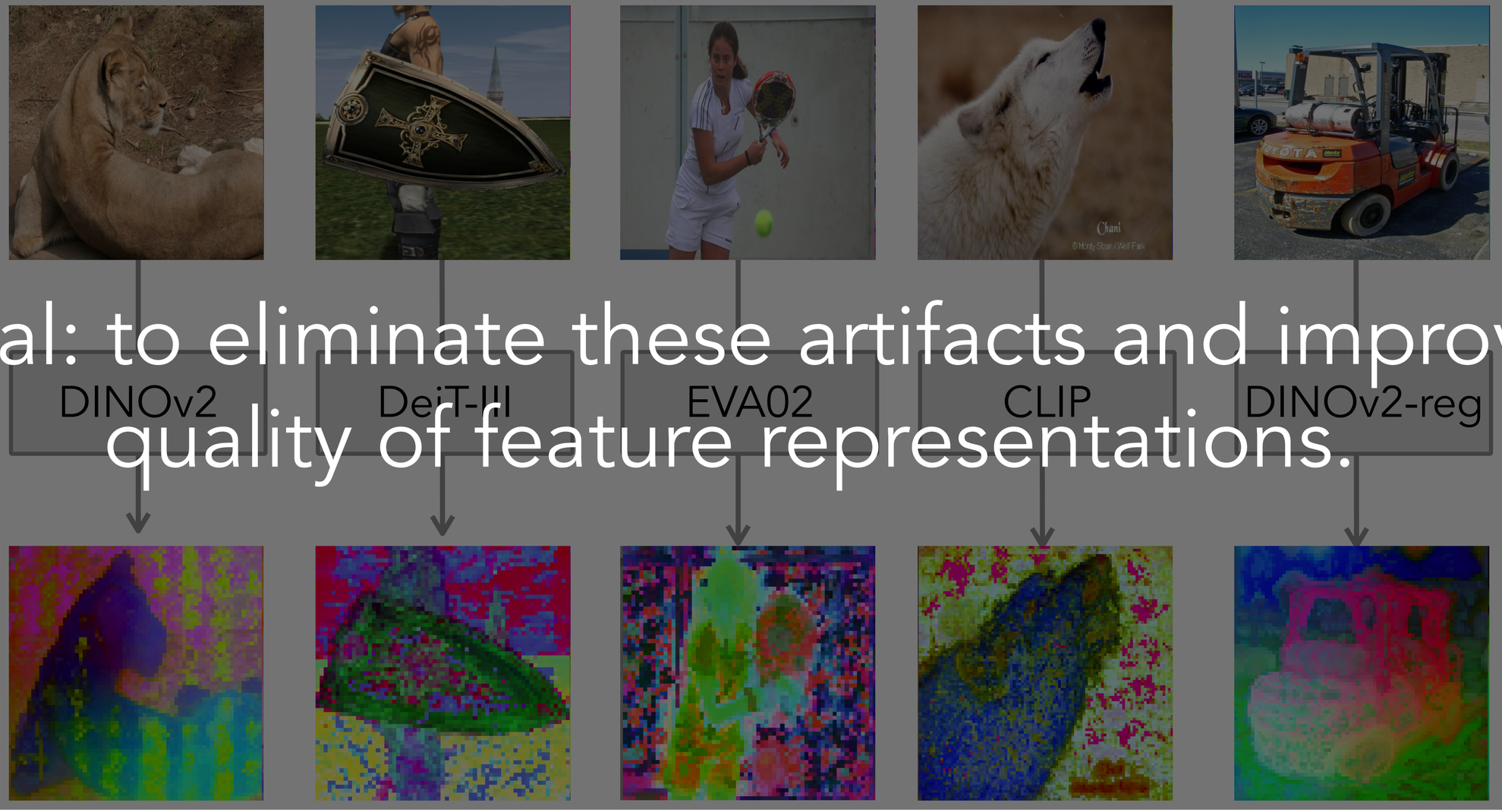
PCA2



Noisy PCAs

Noisy Feature Norms

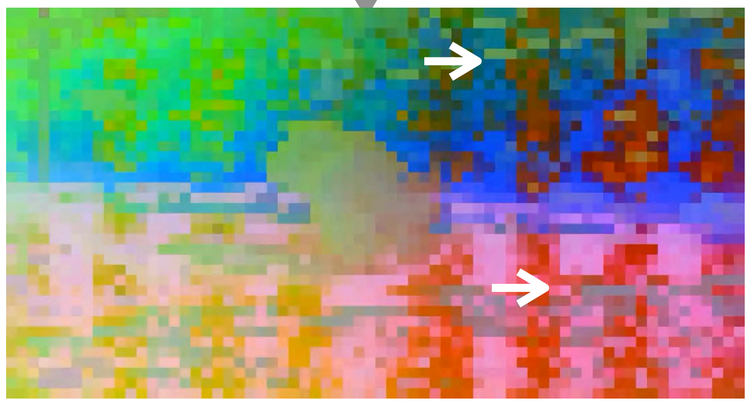
Artifacts in different ViTs



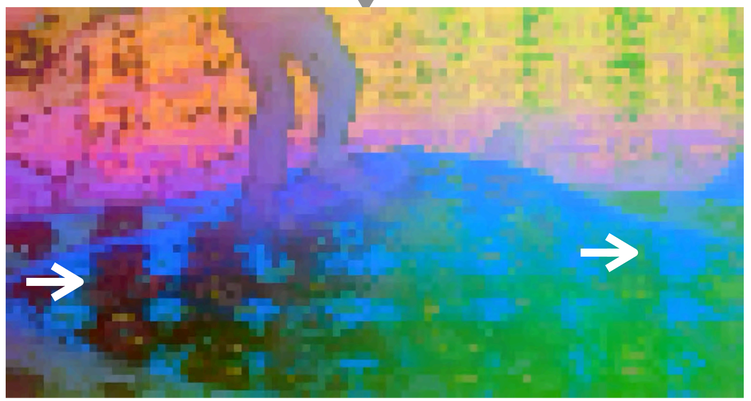
Goal: to eliminate these artifacts and improve quality of feature representations.

Where do these artifacts come from?

DINOv2

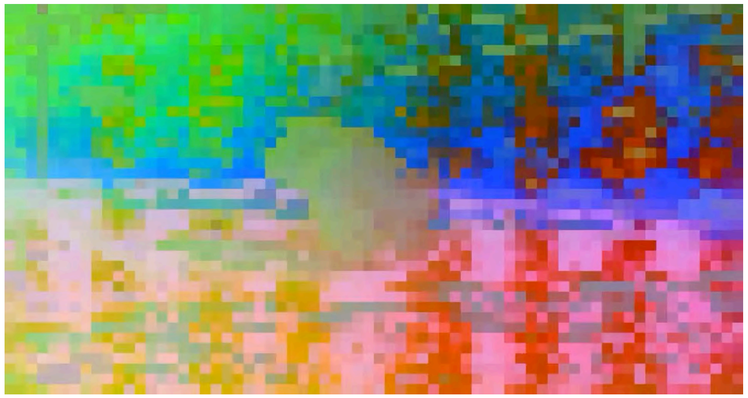


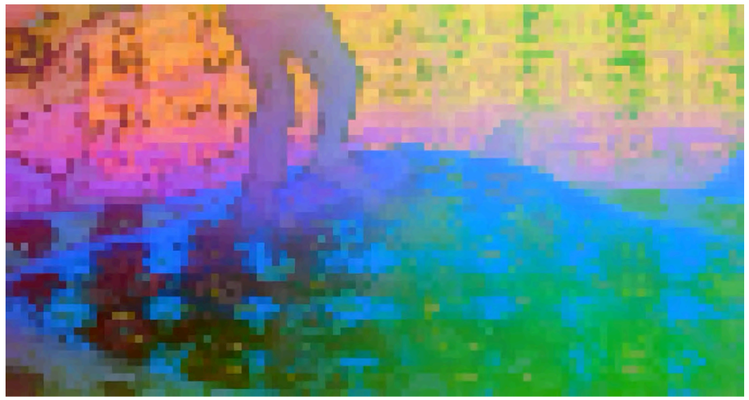
DINOv2



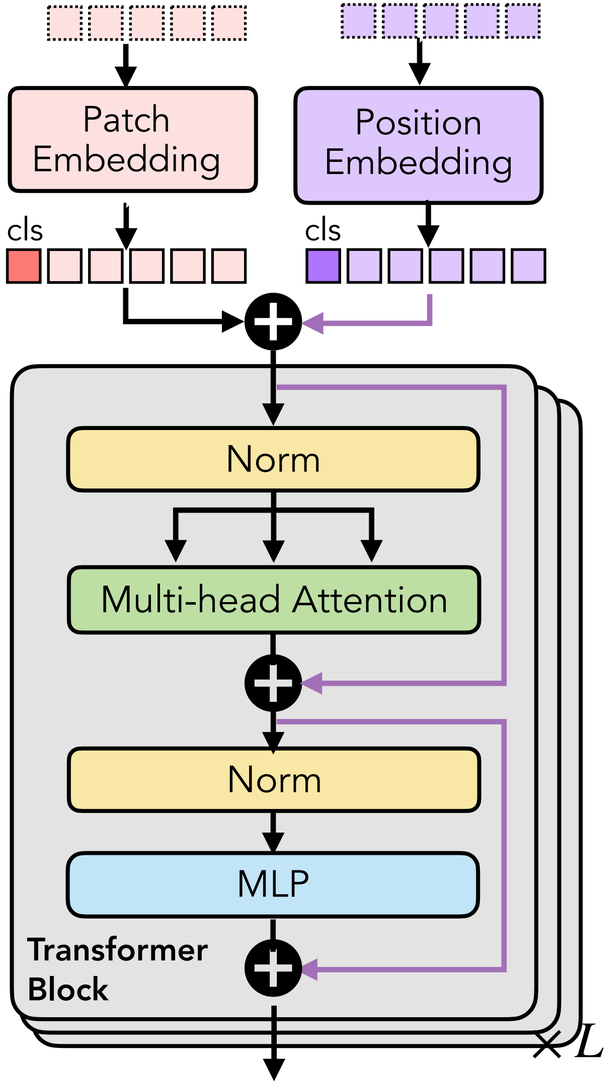
Observation: Although image content changes, artifacts remain nearly fixed in their absolute positions

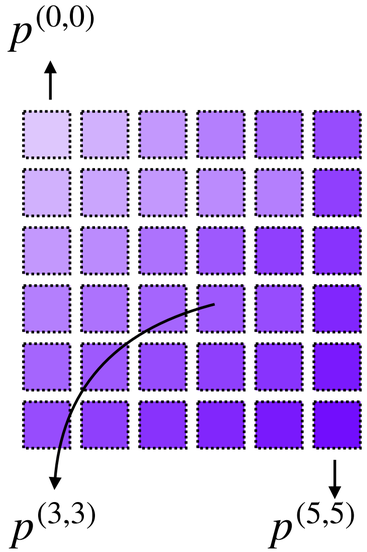
Where do these artifacts come from?





Positional embeddings are applied consistently across images, regardless of content.





$$p^{(i,j)} = p^{(i)} \oplus p^{(j)}$$

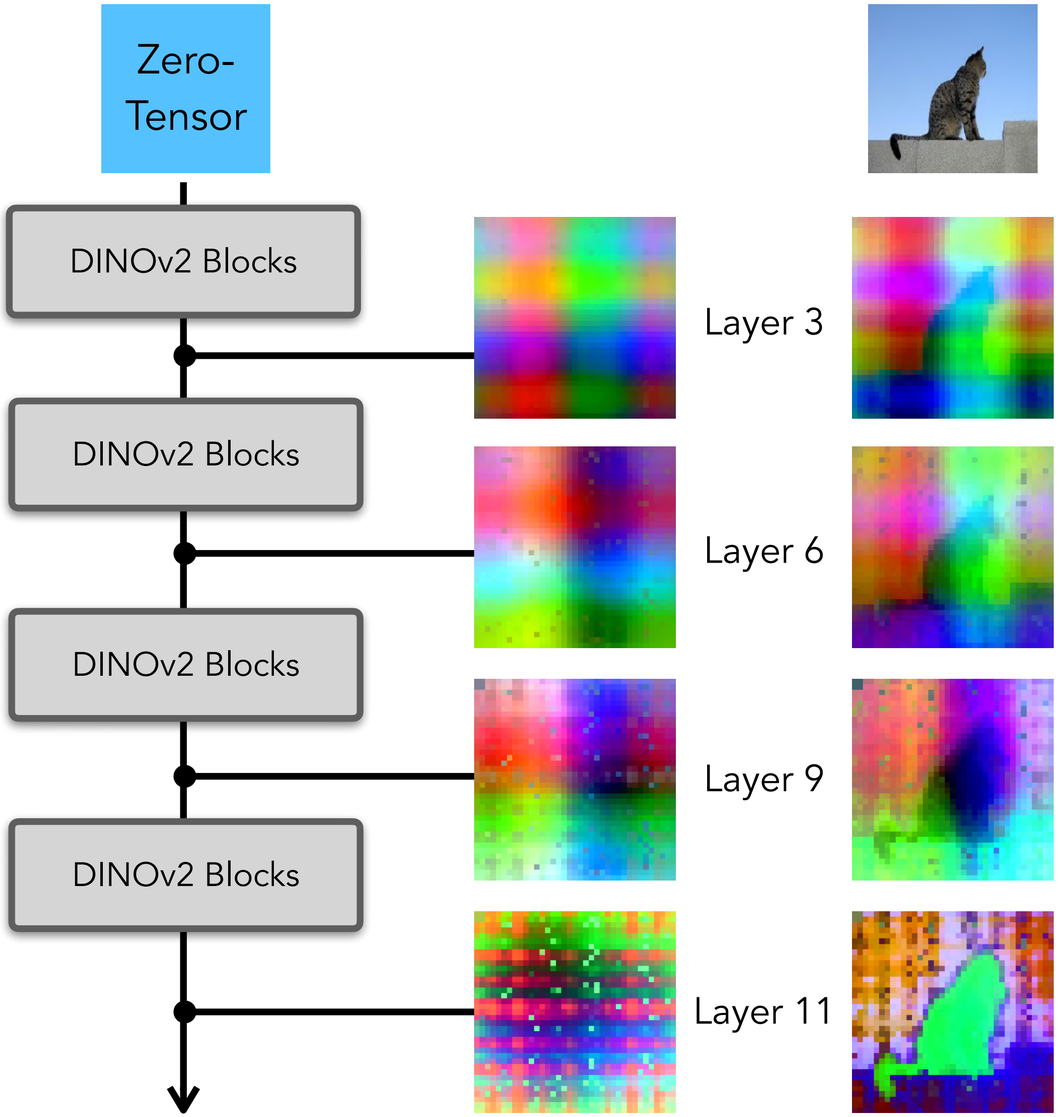
$$p^{(i)}_{2k} = \sin\left(\frac{i}{10000^{2k/D}}\right)$$

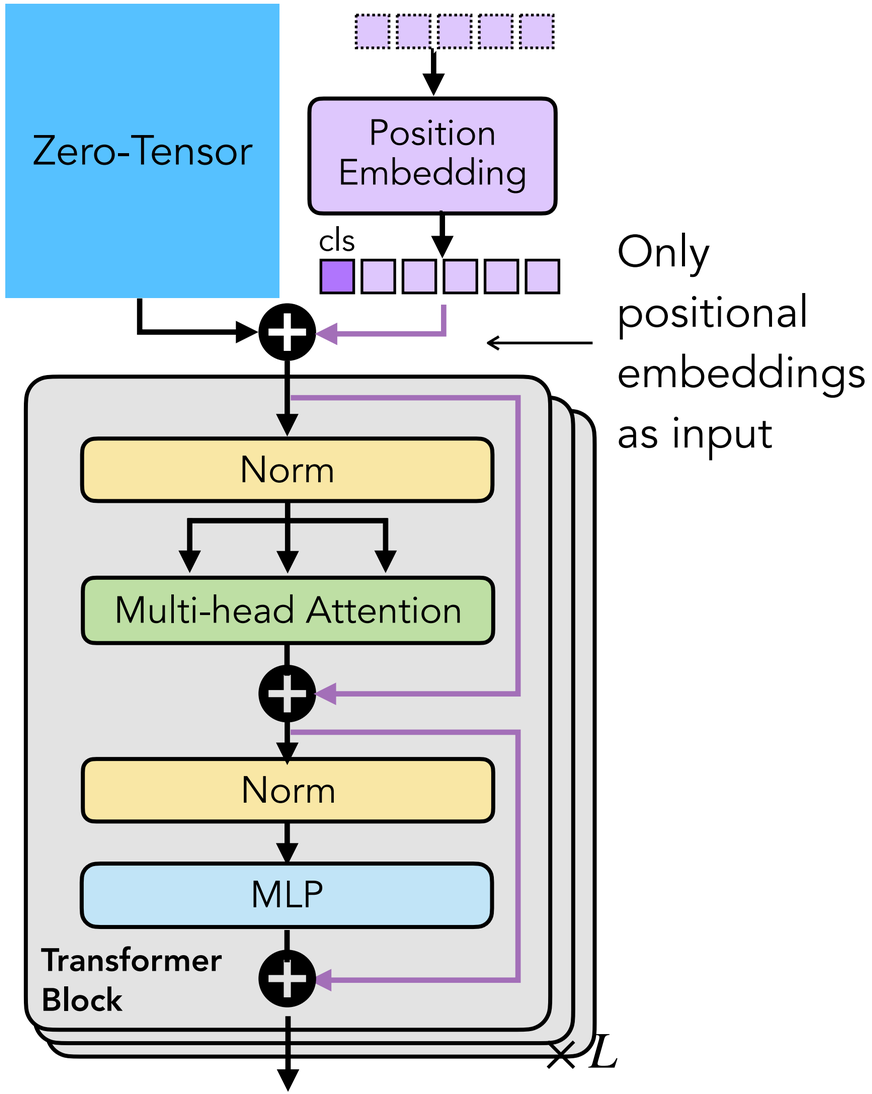
$$p^{(i)}_{2k+1} = \cos\left(\frac{i}{10000^{2k/D}}\right)$$

$$p^{(j)}_{2k} = \sin\left(\frac{j}{10000^{2k/D}}\right)$$

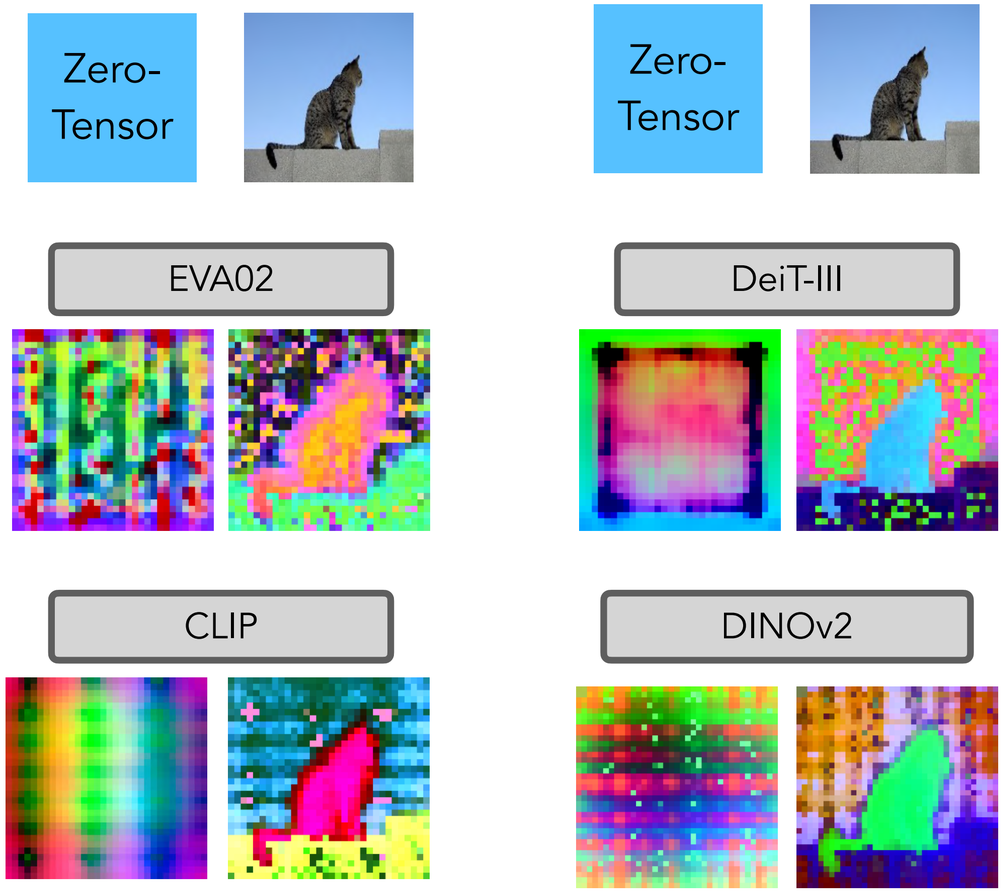
$$p^{(j)}_{2k+1} = \cos\left(\frac{j}{10000^{2k/D}}\right)$$

Test the hypothesis: input a zero-tensor

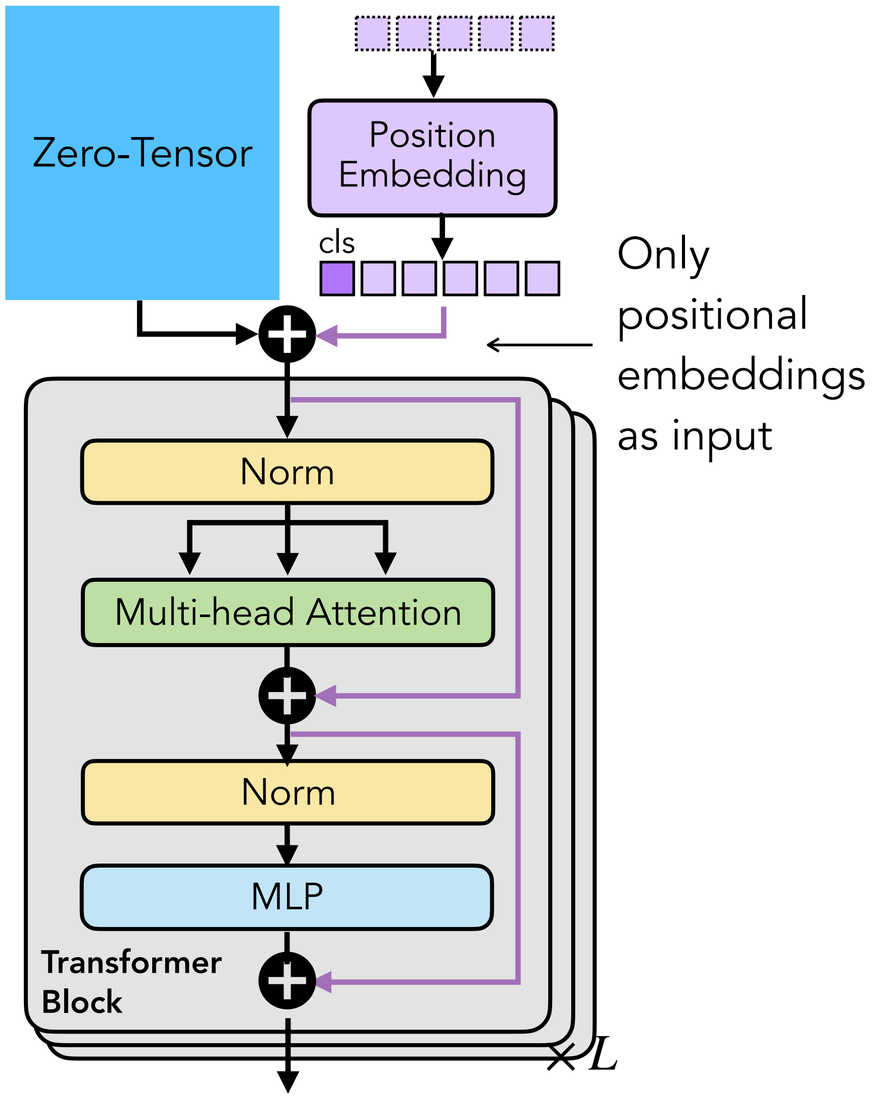




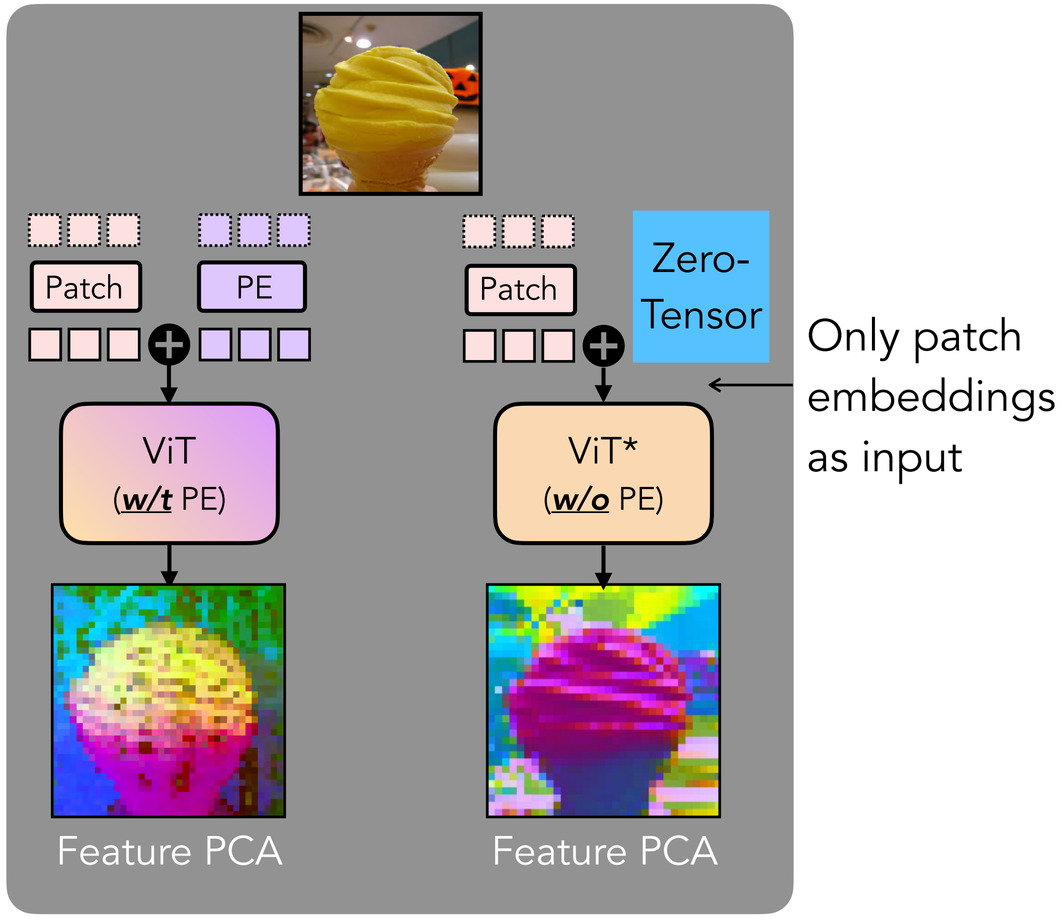
Test the hypothesis: input a zero-tensor



Inputting zero tensors into different ViTs reveals artifacts.

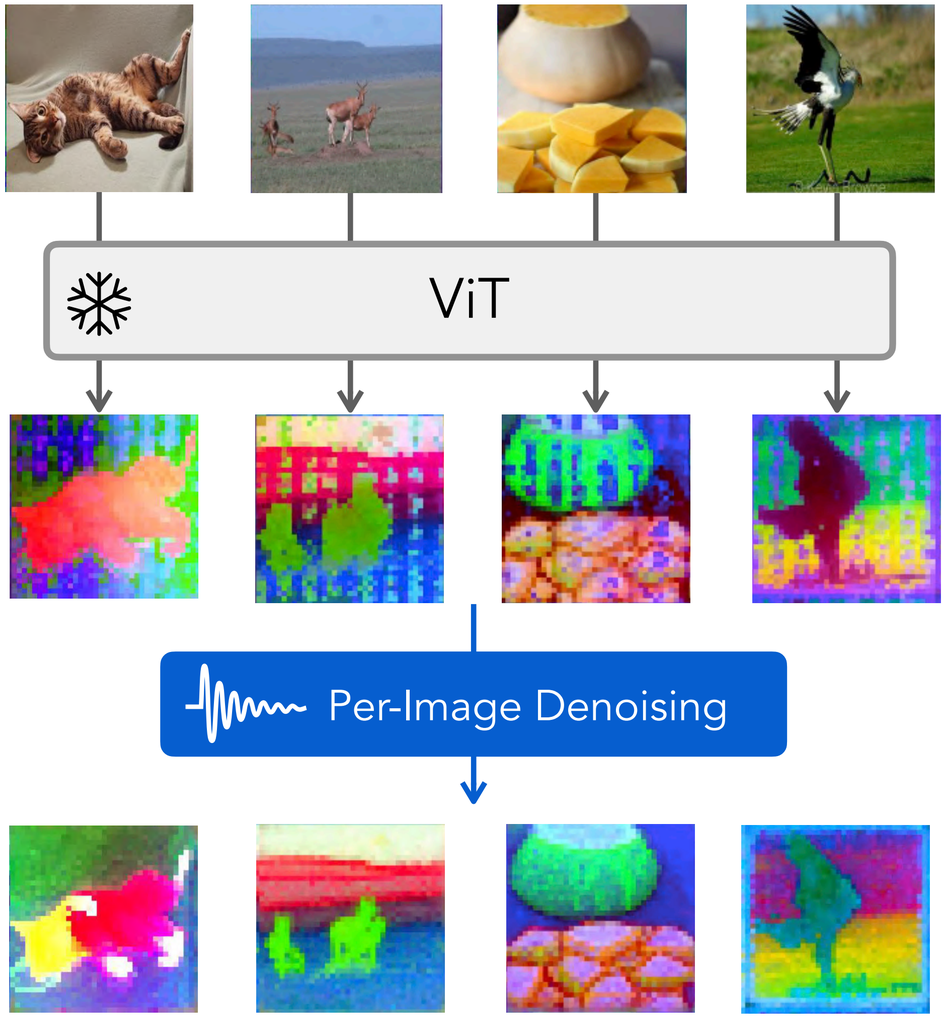


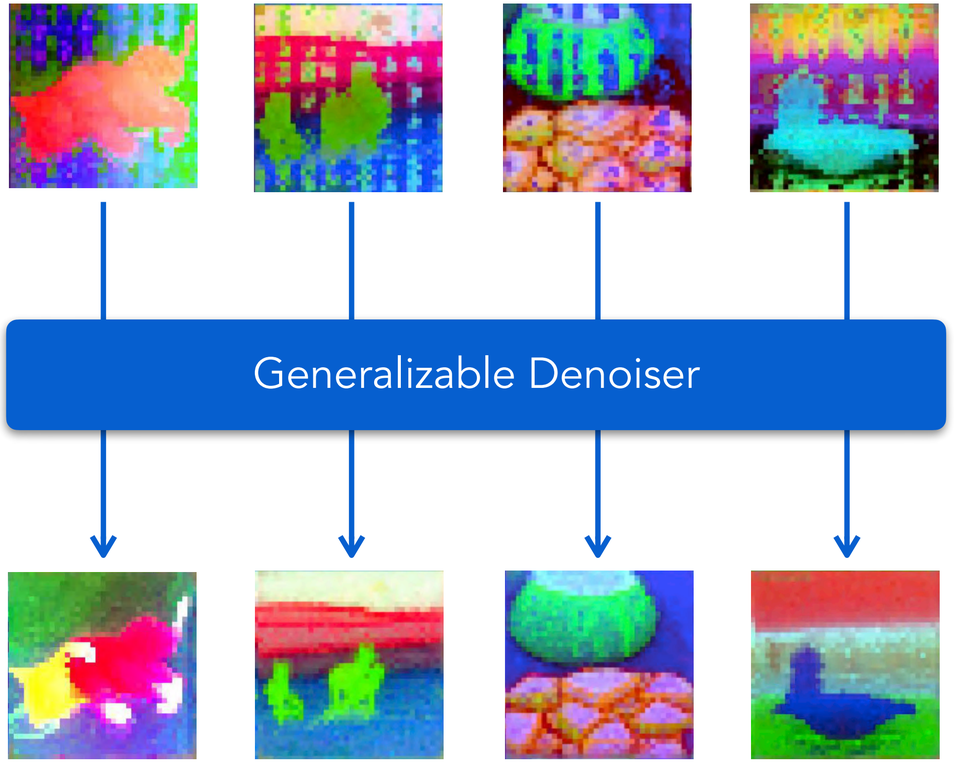
Test the hypothesis: remove PEs



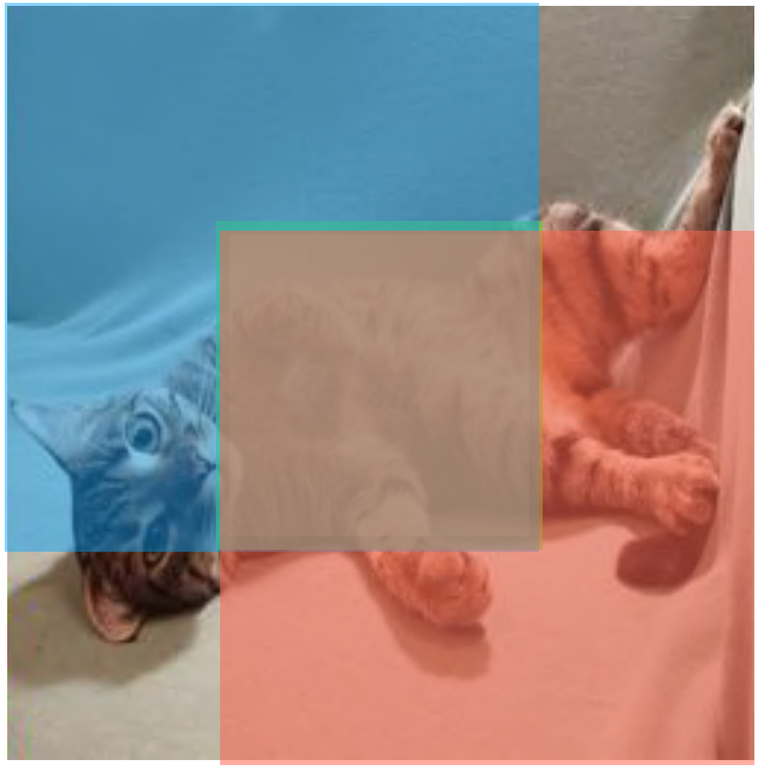
Remove PE -> Artifacts disappear

Denoising Vision Transformers

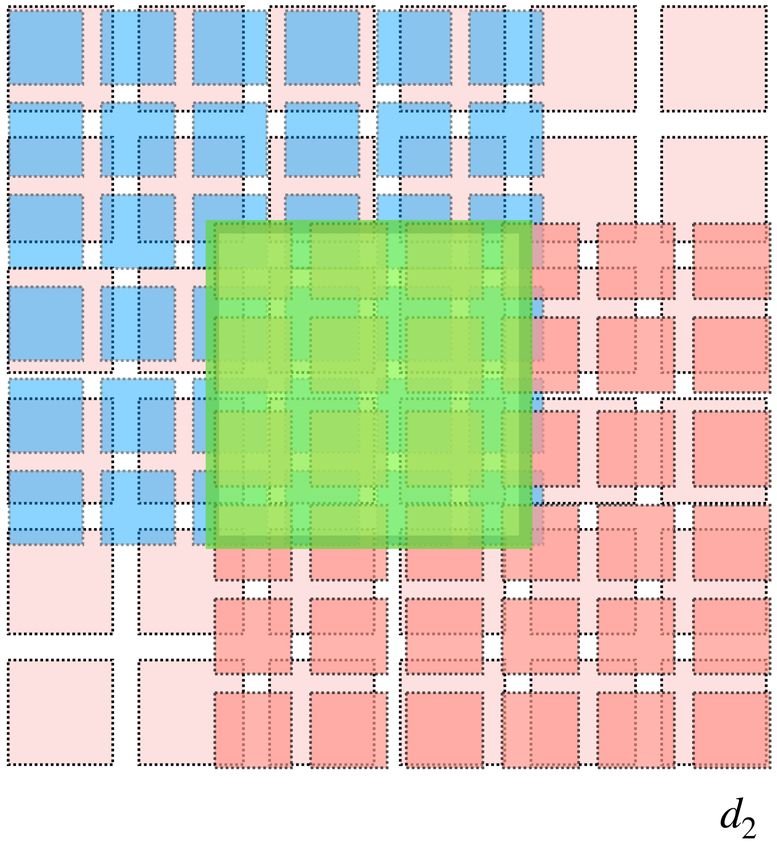




[ECCV 2024] “Denoising Vision Transformers.” Yang et al.



DINOv2



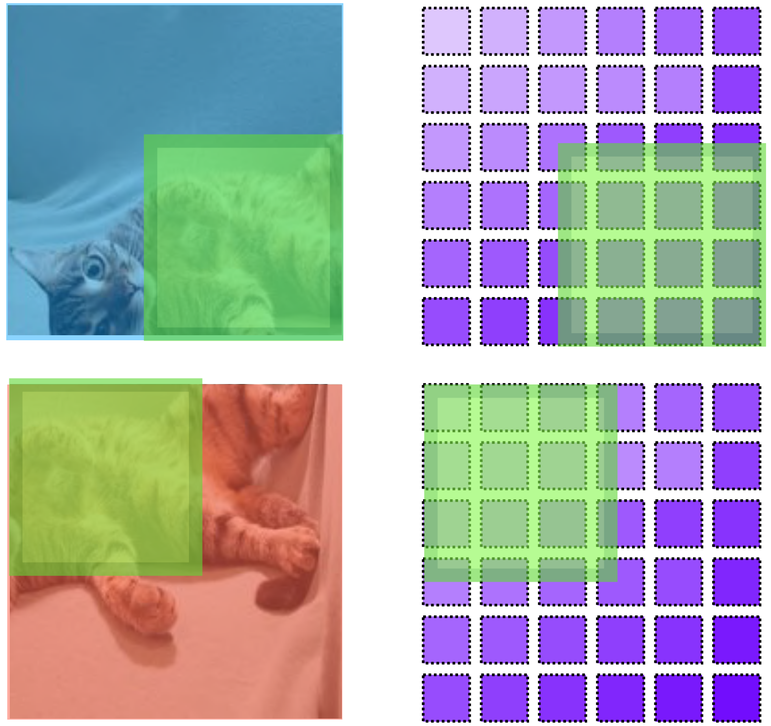
$$x_1 = \text{crop}(x, \theta_1),\ x_2 = \text{crop}(x, \theta_2)$$

$$d_1 = \text{ViTs}(x_1) \approx f(x_1) + g(E_{pos}(x_1))$$
$$d_2 = \text{ViTs}(x_2) \approx f(x_2) + g(E_{pos}(x_2))$$

Feature & artifact consistency

View 1

View 2

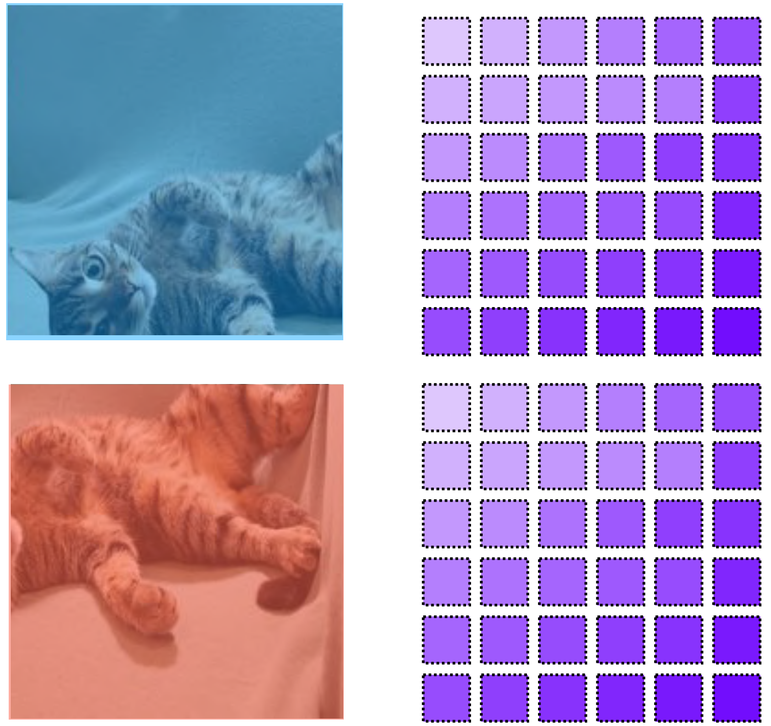


Semantic feature consistency:

Green region should be the same

View 1

View 2



Artifact consistency:

Purple region should be the same

Modeling consistency with neural fields

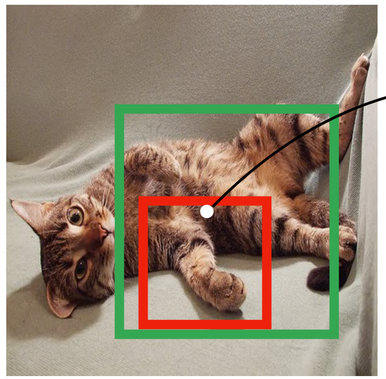


Image Pixels

$(i,j) \longrightarrow f(i,j)$ remains the same

Neural fields $f_\phi : \mathbb{R}^n \rightarrow \mathbb{R}^m$,

ϕ is usually a MLP.

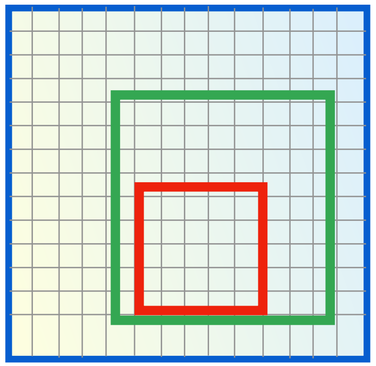
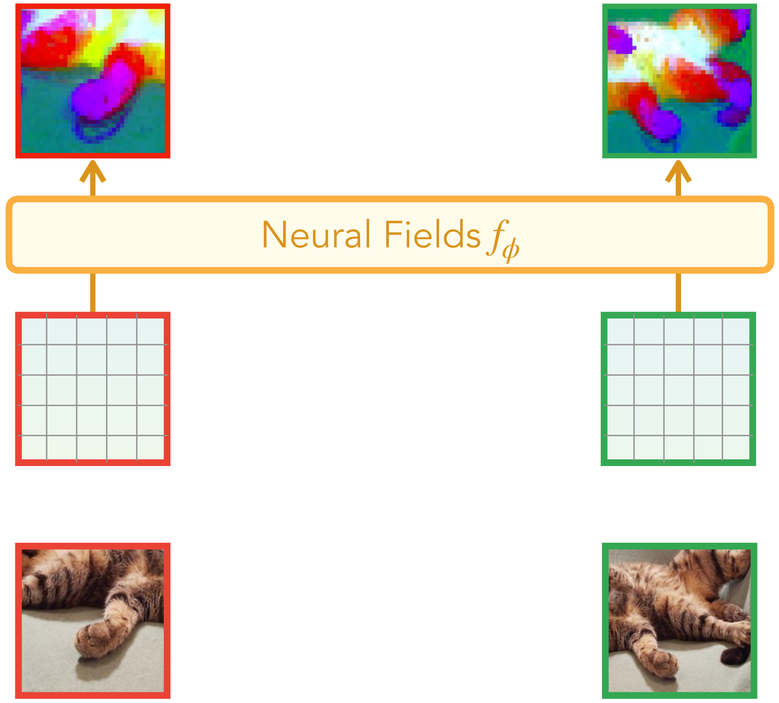
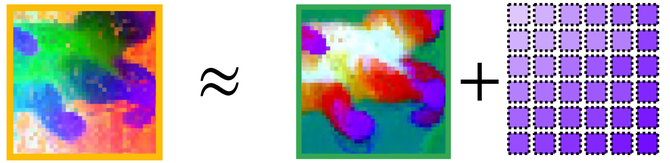


Image Coordinates



$$d_1 = \text{ViTs}(x_1) \approx f(x_1) + g(E_{pos}(x_1))$$

$$d_2 = \text{ViTs}(x_2) \approx f(x_2) + g(E_{pos}(x_2))$$



Clean features -> neural fields

Artifacts -> purple grids

Modeling consistency with neural fields

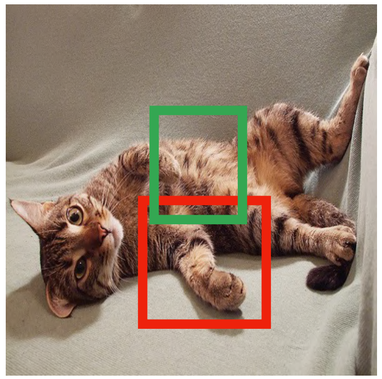
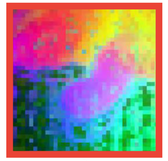
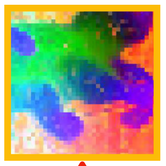


Image Pixels

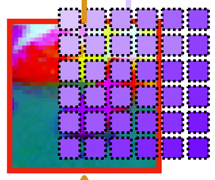
ViT

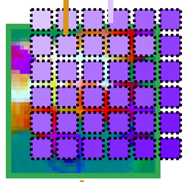




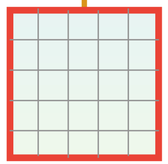
+

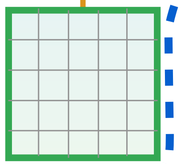
+





Neural Fields $\mathcal{F}$





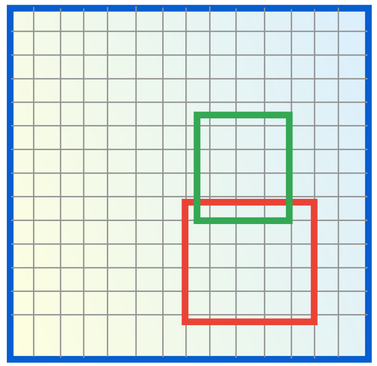
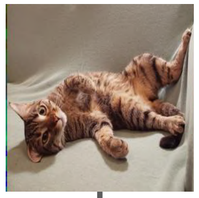
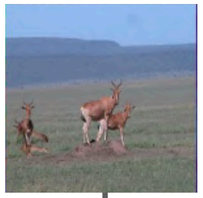
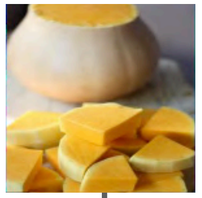


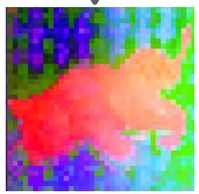
Image Coordinates

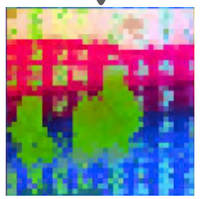


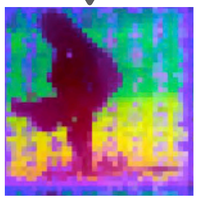




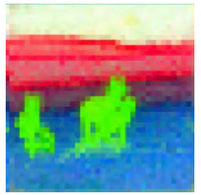
ViT

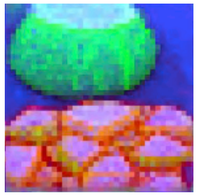


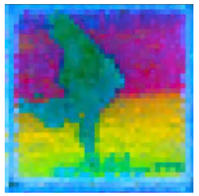




Per-Image Denoising

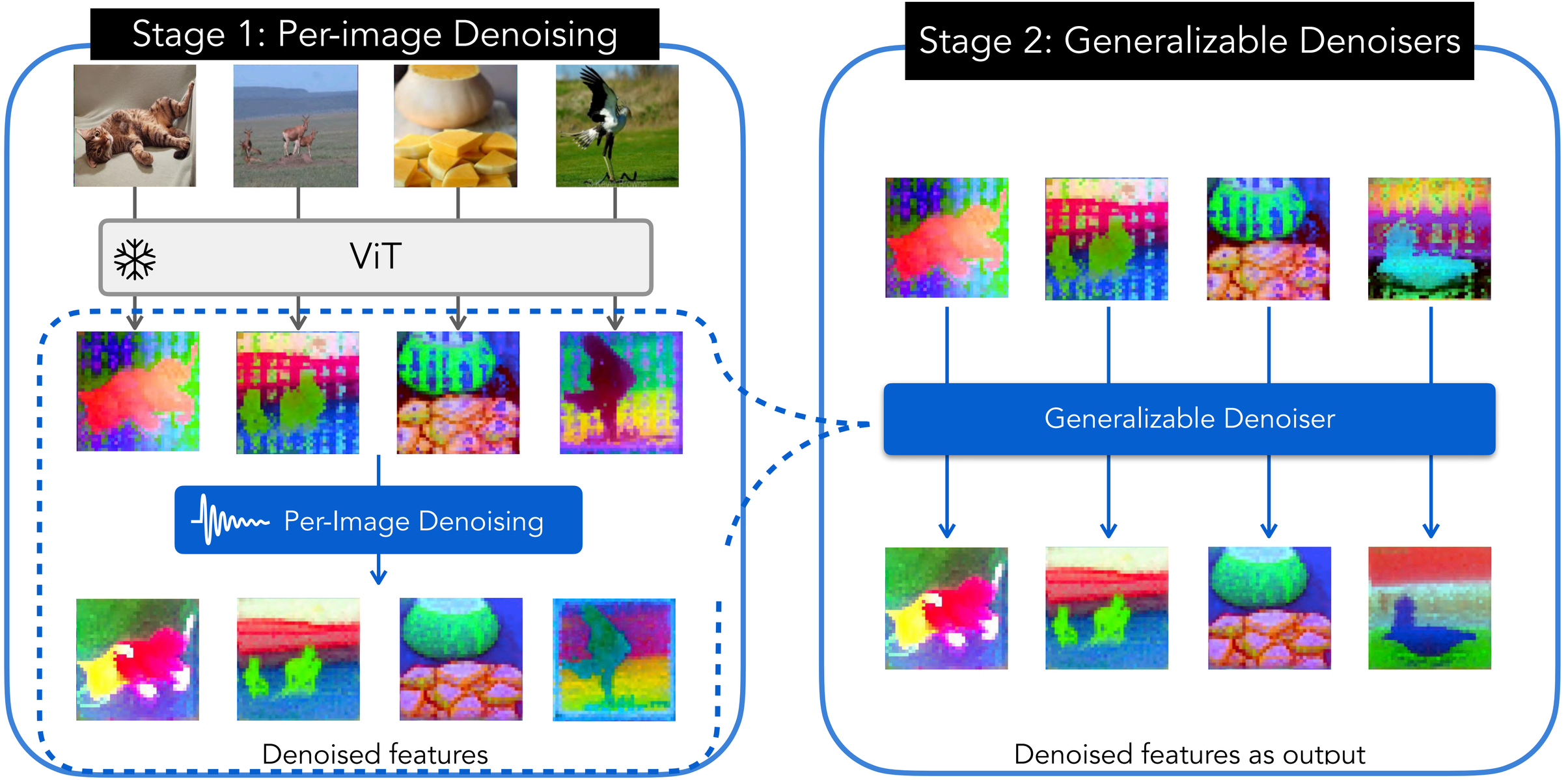






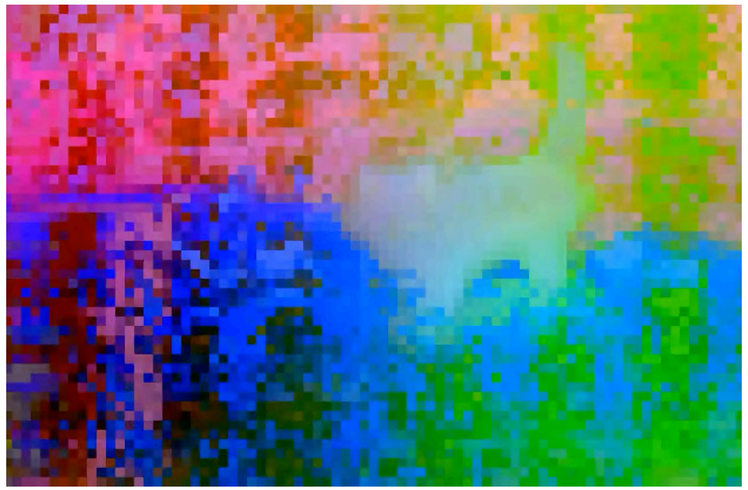
Denoised features

Generalizable feature denoisers

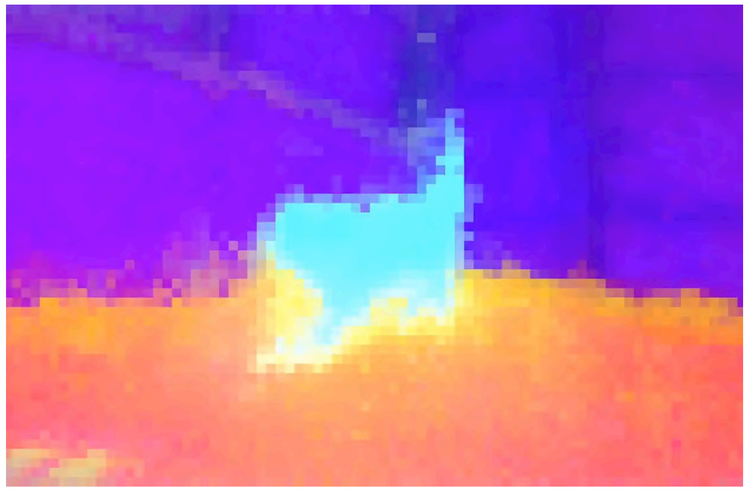


Denoised features

Input

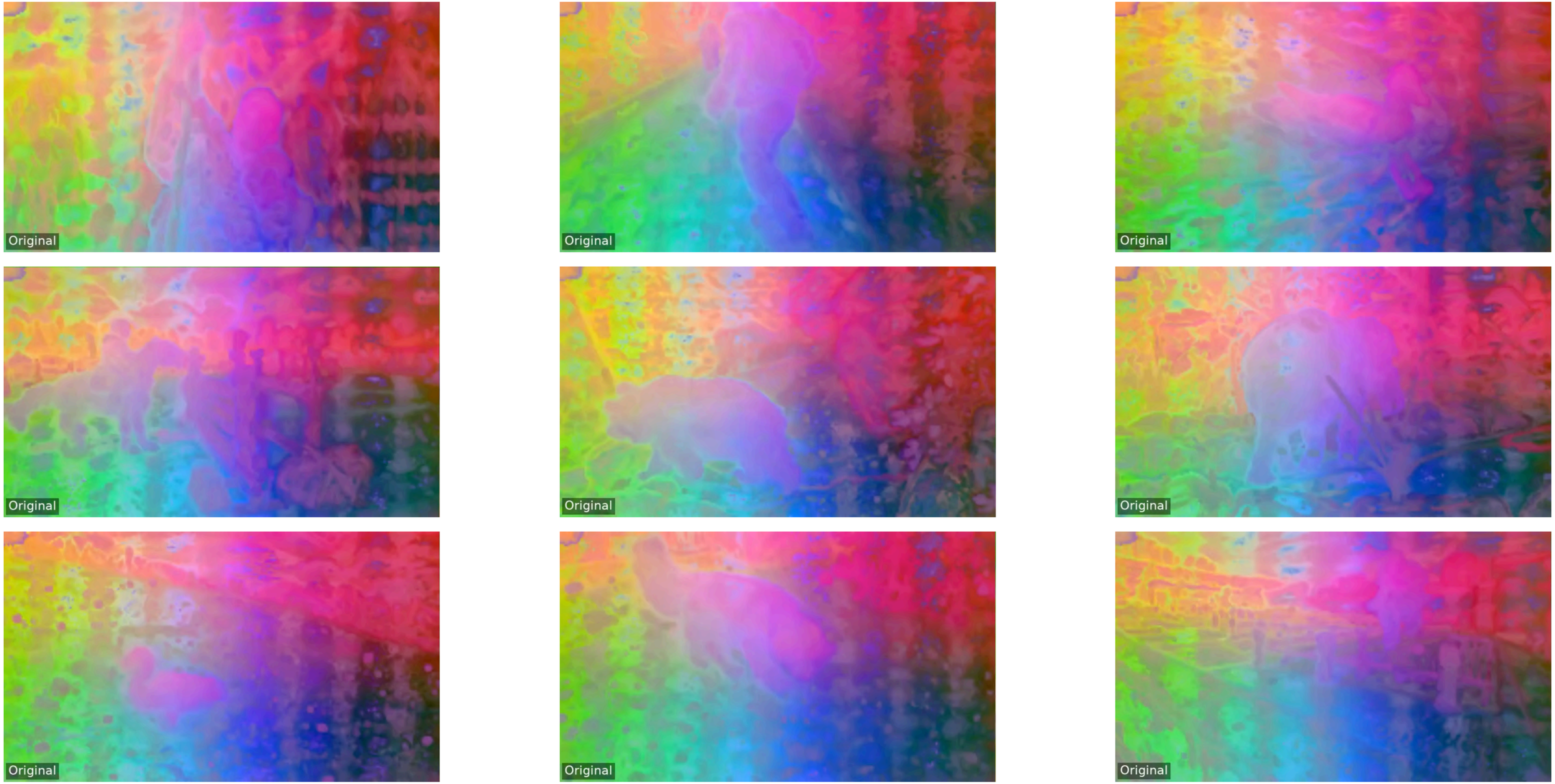


Original Features

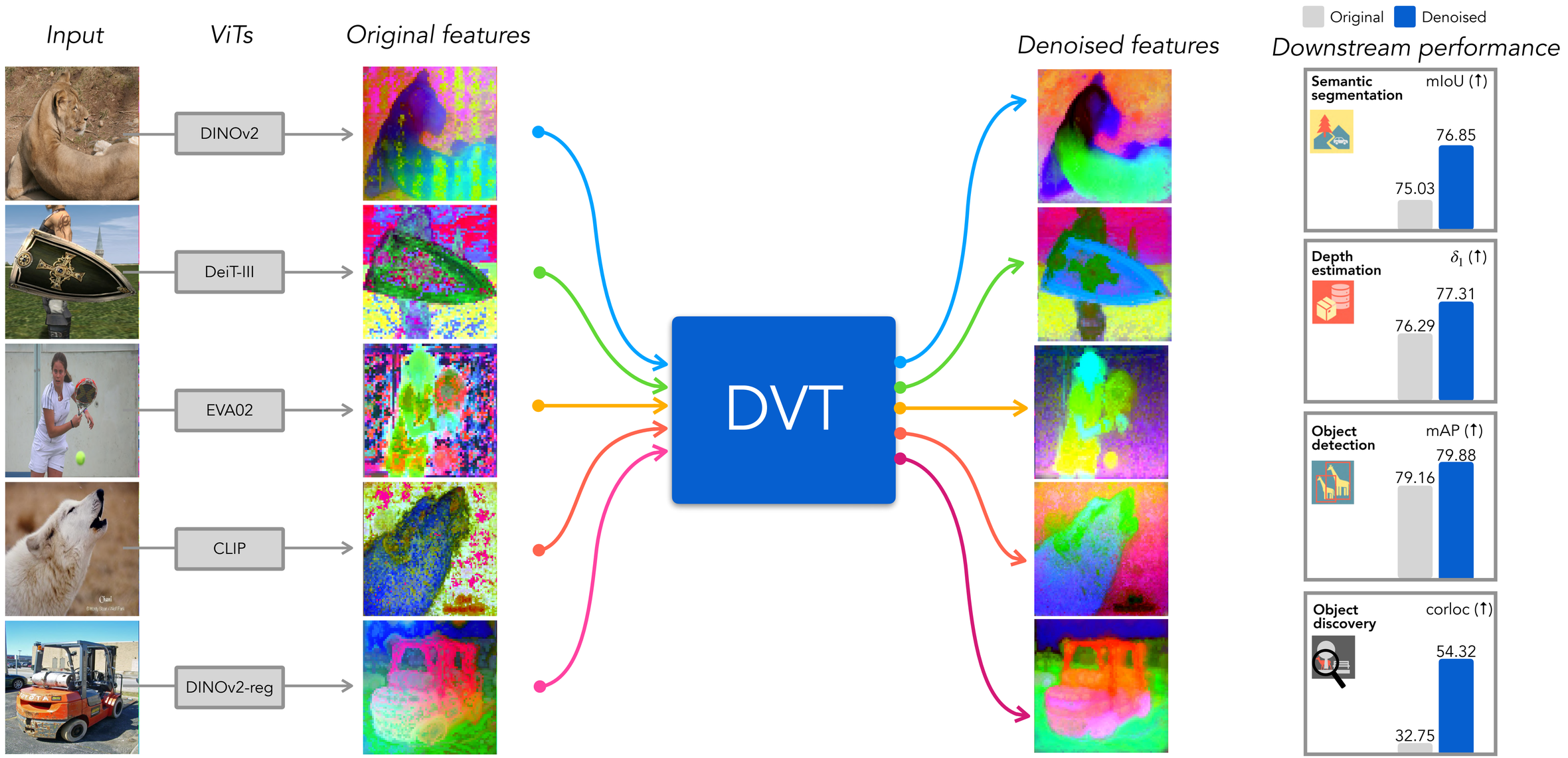


DVT Features

Denoised features



Removing artifacts leads to improved performance, for **free**



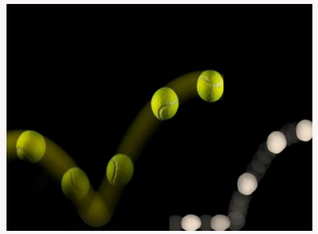
Do foundation models understand physics?

Common VQA

Q What are the things I should be cautious about when I visit here?

A When visiting the pier over the lake, there are a few things you should be cautious about. First, ensure that you …

Physical Object Property

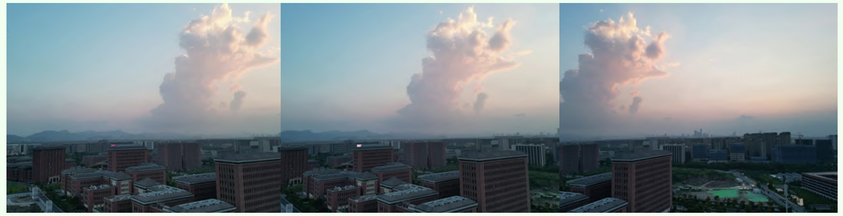


Q Which object has greater elasticity?
A. Green ball ✓ B. White ball
C. Same elasticity
D. Cannot determine

Physical Object Relationships

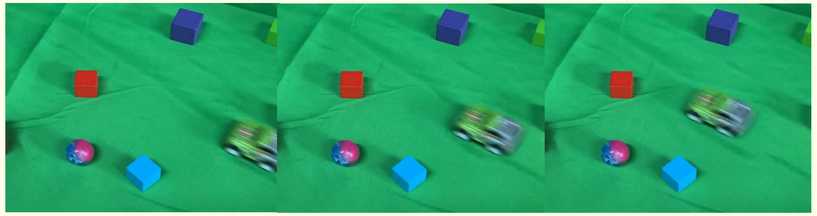
Q What is the object closest to the teacup in the Figure?
A. The pastry B. The peach
C. The knife D. The spoon ✓

Physical Scene Understanding



Q How does the viewpoint alter?
A. Moves downward B. Rotates to the right ✓
C. Moves upward D. Rotates upward

Physics-based Dynamics



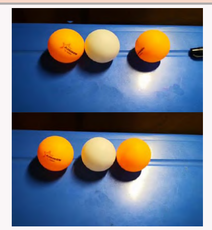
Q Which object will the cart hit first?
A. The red cube ✓ B. The blue cube
C. The green cube D. The yellow cube

[ICLR 2025] "PhysBench: Benchmarking and Enhancing Vision-Language Models for Physical World Understanding." Zhou et al.

Question categories

Physical Object Property

Attribute
Q Given that the applied force is the same, which object in the images has higher stiffness?
A The object in the first image.



Mass
Q What is the mass relationship between the three ping-pong balls?
A The mass of the three ping-pong balls is identical.

Number
Q Which color of balls has the largest number?
A Blue balls.

Color
Q What is the color of the leftmost spectrum? **A** Red.

Physical Object Relationships

Location
Q What is beneath the egg? **A** Mushrooms.

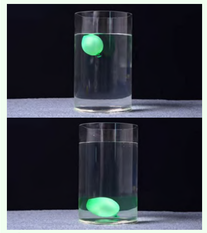
Distance
Q What is the distance between the yellow cube and the blue ball? (The blue cube has a width of 2 cm.) **A** About 7cm.

Depth
Q Which marked object is closest to the camera?
A Option B.

Size
Q What is the color of the largest cube? **A** Red.

Velocity
Q Which car has a higher average speed?
A The red one.

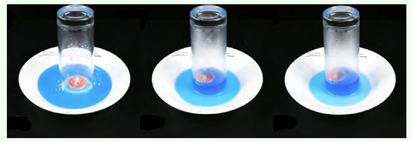
Physical Scene Understanding



Temperature
Q Is the phenomenon observed in the video caused by adding cold water or hot water?
A Hot water.

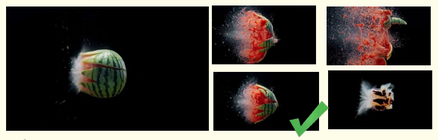
Viewpoint
Q How does the focal length of the camera change?
A The focal length increases.

Light
Q How might the light source in the image have changed?
A It appears to have shifted from the left side of the image to the right side.



Air Pressure
Q What causes the change in water level in the cup?
A The combustion lowers the air pressure in the cup.

Physical-based Dynamics



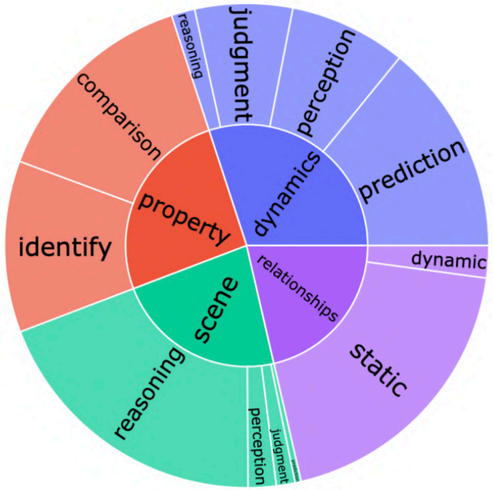
Collision
Q Which scene, depicted in the images, occurs first?

Throwing
Q Which can is the ball most likely to land in?
A The white can.

Figure 1 Figure 2 Figure 3

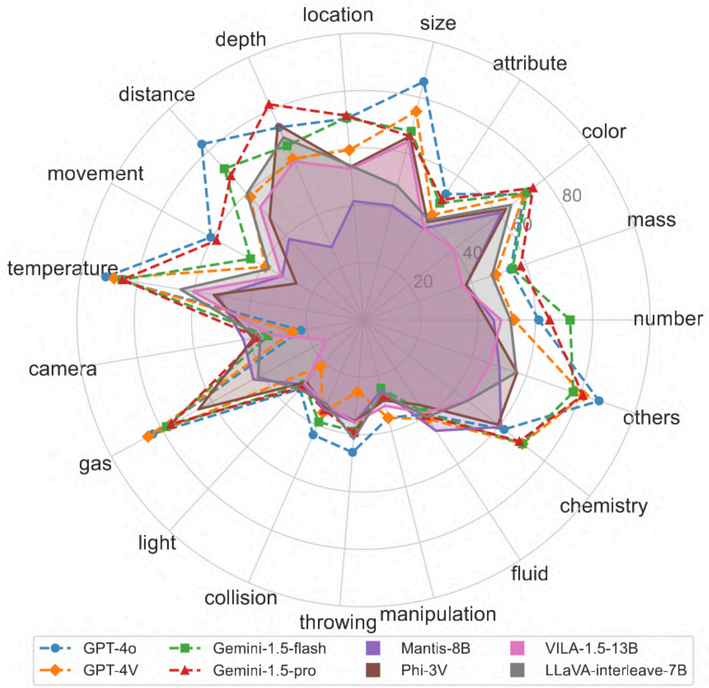
Manipulation
Q What is the correct sequence of images to make a gift box containing the perfume bottle?
A first, image 1, followed by image 3, and finally image 2.

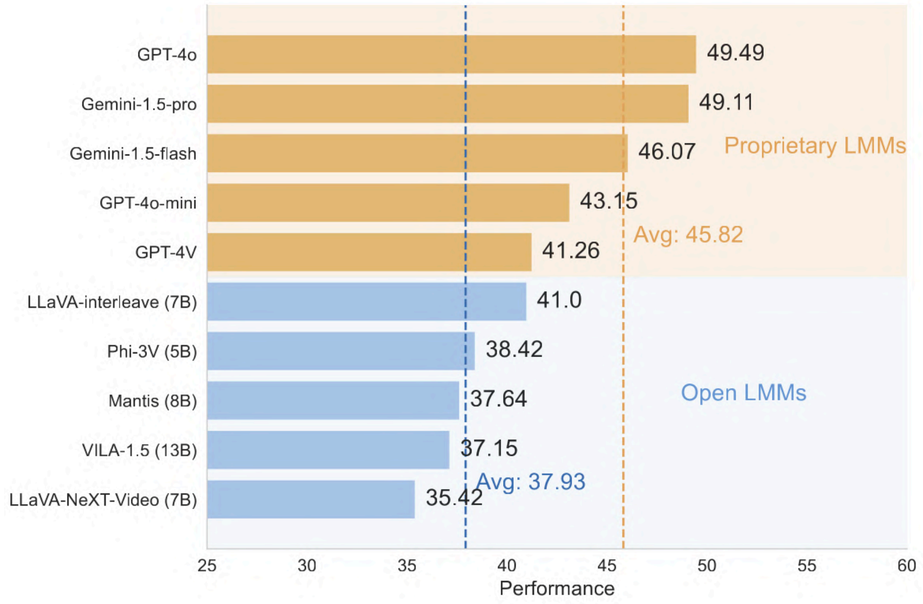
Fluid
Q Which object has the lowest viscosity?
A The white liquid.



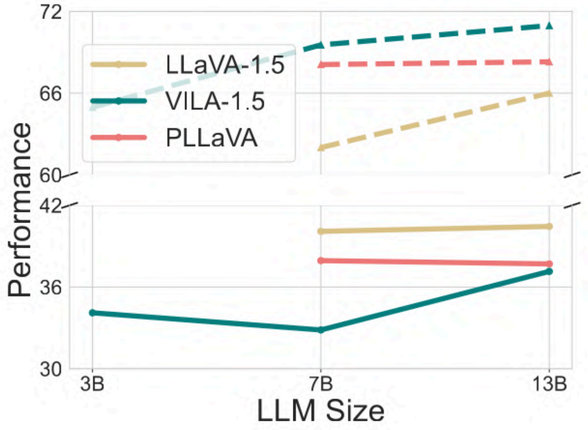
Statistic	Number
Total questions	10,002
- only one image	1,766 (18.6%)
- only one video	2,749 (44.8%)
- interleave	1,902 (20.1%)
Unique number of images	10,058
Unique number of videos	3,260
3D Assets	678
Maximum question length	48
Maximum choice length	20
Average question length	16.5
Average choice length	4.4

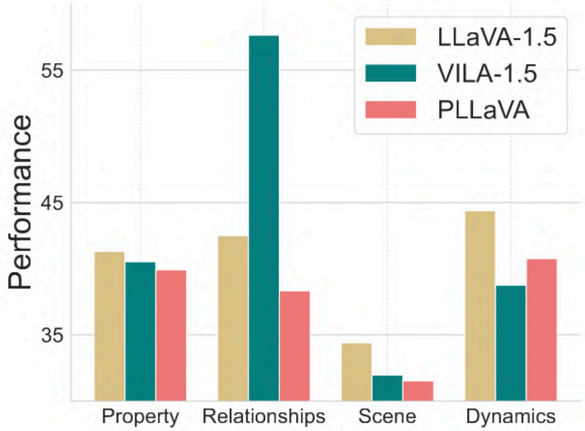
	Size	Format	Property	Relationships	Scene	Dynamics	Avg
Random Choice	-	-	25.00	25.00	25.00	25.00	25.00
Human	-	-	97.10	95.67	94.91	95.68	95.87
Image VLM							
InstructBLIP-t5-xl Dai et al. (2024)	4B	merge	35.35	36.67	37.45	35.95	36.24
InstructBLIP-t5-xxl Dai et al. (2024)	12B	merge	41.11	38.47	37.89	36.42	38.51
InstructBLIP-7B Dai et al. (2024)	7B	merge	21.94	29.00	19.53	27.45	23.82
InstructBLIP-13B Dai et al. (2024)	13B	merge	31.69	33.19	23.13	30.64	29.94
BLIP-2 Li et al. (2023c)	12B	merge	41.70	40.83	36.25	36.93	38.61
LLaVA-1.5-7B Liu et al. (2023a)	7B	merge	38.44	41.53	**38.60**	42.69	40.09
LLaVA-1.5-13B Liu et al. (2023a)	13B	merge	41.31	42.50	34.40	44.38	40.45
LLaVA1.6-mistral Liu et al. (2024b)	7B	merge	29.77	22.22	8.54	20.58	20.30
LLaVA1.6-vicuna Liu et al. (2024b)	7B	merge	40.26	59.72	**38.60**	42.65	42.28
Qwen-VL-Chat Bai et al. (2023b)	9B	merge	35.97	43.33	26.47	41.27	35.63
InternVL-Chat1.5 Chen et al. (2024c)	26B	merge	**53.08**	**70.14**	37.01	**44.78**	**47.51**
Cambrian-8B Tong et al. (2024)	8B	merge	23.27	17.92	23.02	29.29	24.61
Claude-3-opus Anthropic (2024)	-	merge	41.97	40.97	30.63	36.50	37.00
Claude-3-sonnet Anthropic (2024)	-	merge	37.86	40.00	32.23	36.89	36.18
Claude-3-haiku Anthropic (2024)	-	merge	43.28	53.33	30.06	39.93	39.44
Claude-3.5-sonnet Anthropic (2024)	-	merge	46.46	41.11	27.89	37.60	38.05
Video VLM							
Video-LLaVA Lin et al. (2023a)	7B	seq	36.82	36.11	33.69	40.52	37.04
Chat-Univi-7B Jin et al. (2023)	7B	seq	19.28	20.97	18.86	28.46	22.19
Chat-Univi-13B Jin et al. (2023)	13B	seq	4.30	11.53	15.67	11.47	10.36
PLLaVA-7B Xu et al. (2024)	7B	seq	**38.02**	35.83	**36.34**	39.89	**37.94**
PLLaVA-13B Xu et al. (2024)	13B	seq	**39.91**	**38.33**	31.52	**40.76**	37.70
General VLM + Interleaved data							
LLaVA-interleave Li et al. (2024b)	7B	seq	47.23	44.62	35.64	37.21	41.00
LLaVA-interleave-dpo Li et al. (2024b)	7B	seq	47.97	42.67	33.73	38.78	40.83
VILA-1.5-3B Lin et al. (2023b)	3B	seq	32.40	33.02	34.84	35.78	34.11
VILA-1.5-3B-s2 Lin et al. (2023b)	3B	seq	33.14	30.26	35.72	33.00	33.07
VILA-1.5-8B Lin et al. (2023b)	8B	seq	33.41	29.88	30.85	35.91	32.85
VILA-1.5-13B Lin et al. (2023b)	13B	seq	40.53	40.15	31.96	36.07	37.15
Phi-3V Abdin et al. (2024)	4B	seq	43.67	37.92	34.93	36.92	38.42
LLaVA-NV Zhang et al. (2024b)	7B	seq	38.33	30.83	34.00	37.17	35.42
LLaVA-NV-dpo Zhang et al. (2024b)	7B	seq	38.83	44.31	33.86	37.21	37.43
Mantis-Idefics2 Jiang et al. (2024)	8B	seq	41.97	41.44	29.53	36.56	37.39
Mantis-LLaVA Jiang et al. (2024)	7B	seq	44.48	30.45	36.25	34.73	36.69
Mantis-siglip-llama3 Jiang et al. (2024)	8B	seq	42.47	32.78	**36.83**	37.51	37.64
Mantis-clip-llama3 Jiang et al. (2024)	8B	seq	40.61	35.11	32.45	38.36	36.92
GPT-4V Achiam et al. (2023)	-	seq	49.59	45.77	26.34	42.15	41.26
GPT-4o Achiam et al. (2023)	-	seq	56.91	**64.80**	30.15	**46.99**	**49.49**
GPT-4o-mini Achiam et al. (2023)	-	seq	53.54	44.24	30.59	42.90	43.15
Gemini-1.5-flash Team et al. (2023)	-	seq	**57.41**	52.24	34.32	40.93	46.07
Gemini-1.5-pro Team et al. (2023)	-	seq	57.26	63.61	36.52	41.56	49.11

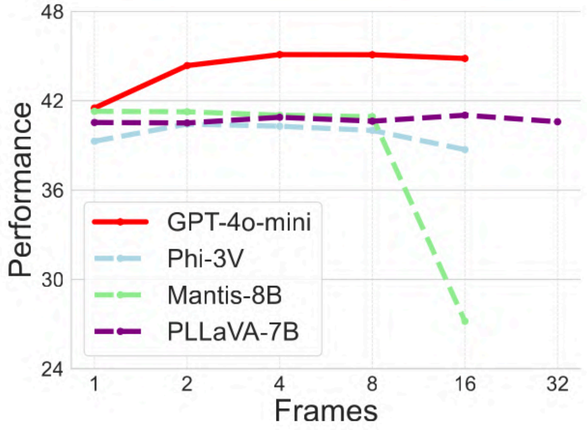




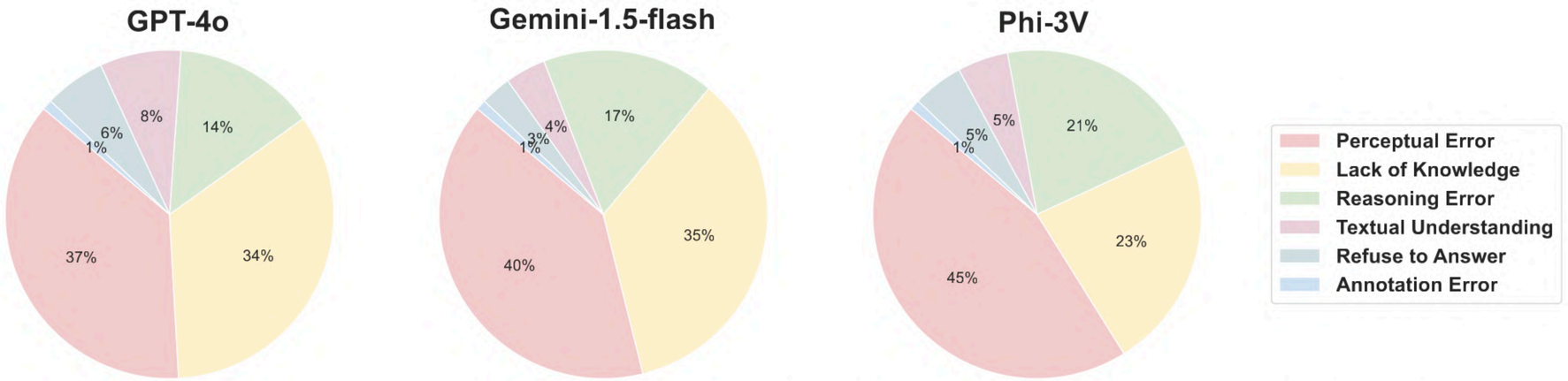
How to improve VLMs' physical understanding?





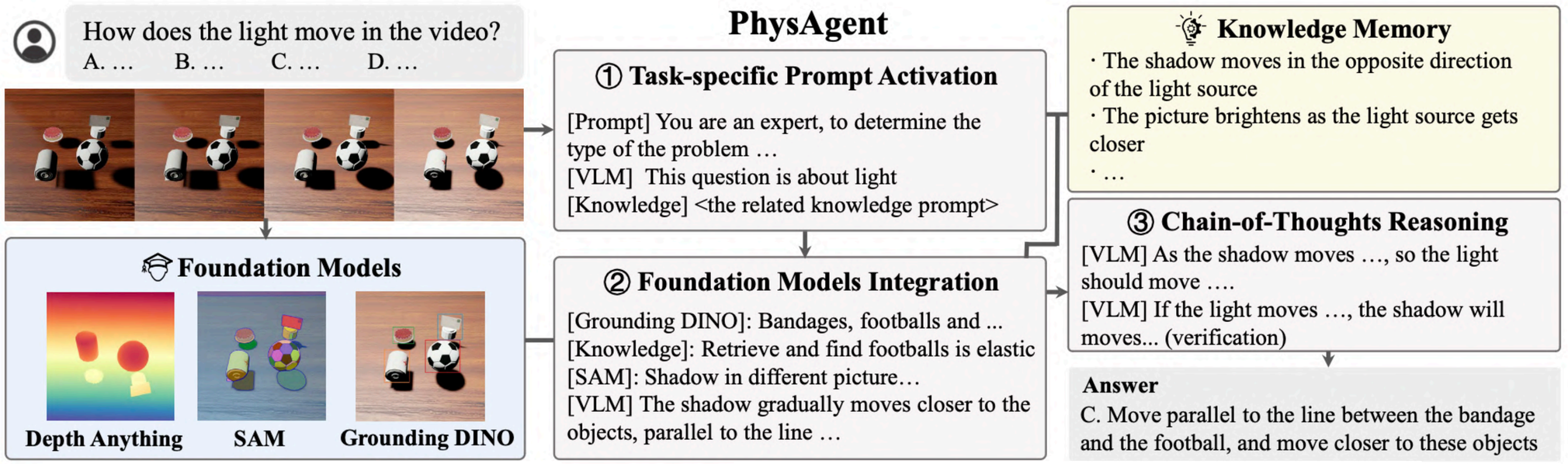


Scaling doesn't help!

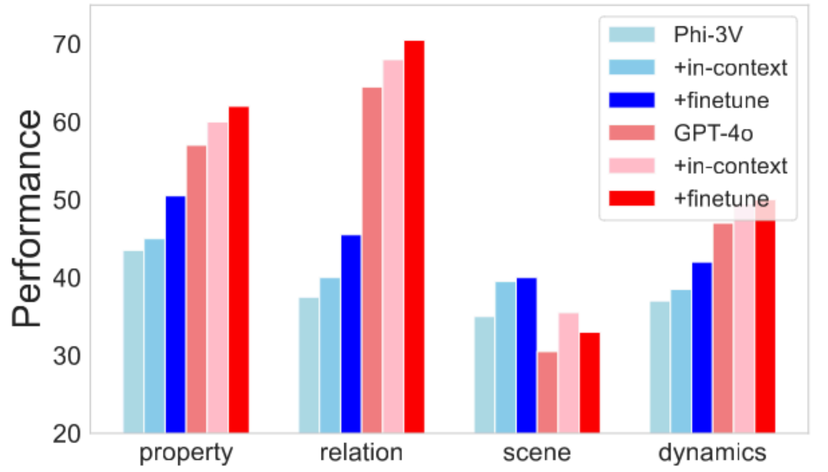


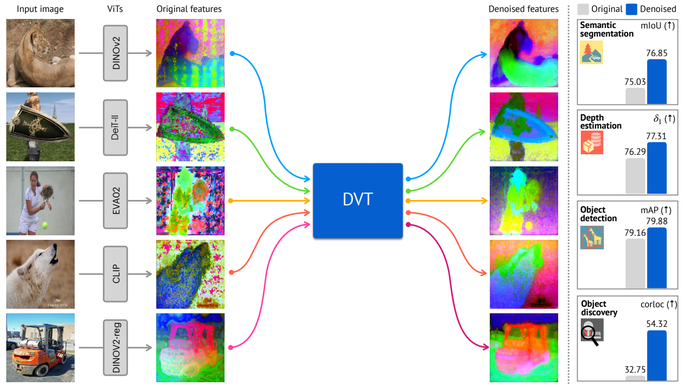
Perceptual errors
& lack of knowledge

PhysAgent: an agentic approach towards better physical understanding



	Property	Spatial	Envir.	Phe.
Phi-3V	38.5	34.4	31.6	32.5
+ CoT	38.8	34.7	31.1	31.9
+ Desp-CoT	25.1	24.1	18.9	21.2
+ PLR	23.1	23.9	19.3	17.1
+ **PhysAgent**	**44.5**	**47.0**	**38.6**	**37.1**
ContPhy	52.1	52.9	37.2	42.8
GPT-4o	53.7	61.7	27.0	34.3
+ CoT	54.5	63.2	26.4	35.1
+ Desp-CoT	51.1	58.8	27.2	32.1
+ PLR	37.8	46.2	15.4	22.1
+ **PhysAgent**	**58.4**	**84.2**	**45.0**	**51.3**

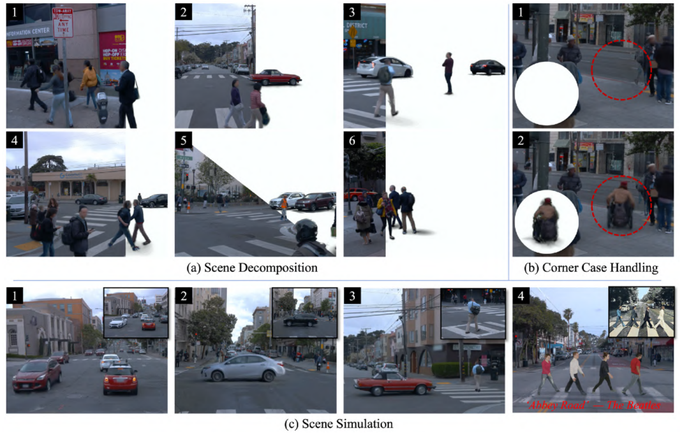


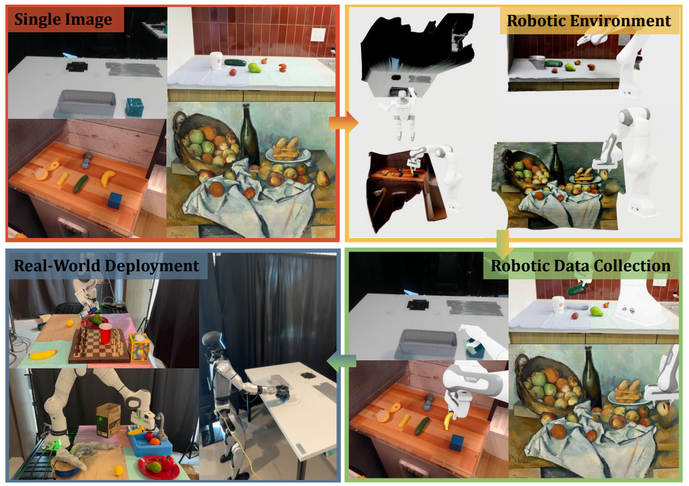


[ECCV 2024] "Denoising Vision Transformers." Yang et al.

[ICLR 2025] "PhysBench: Benchmarking and Enhancing Vision-Language Models for Physical World Understanding." Zhou et al.

Foundation models for vision and physics.

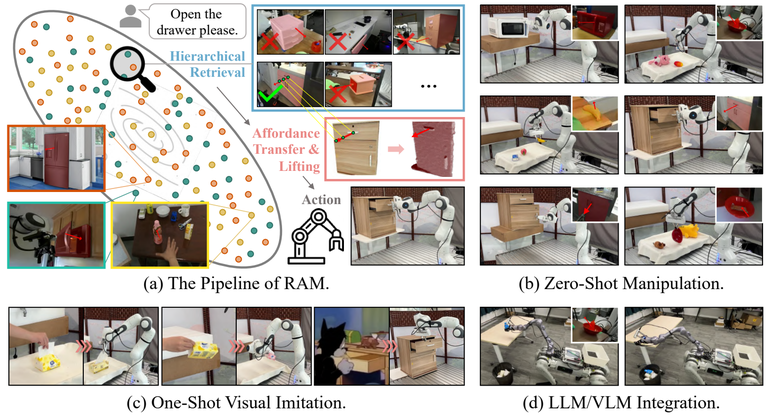




[ICLR 2025] "Omni Urban Scene Reconstruction." Chen et al.

[In submission 2025] "Robot Learning from Any Images" Zhao et al.

3D reconstruction for physical simulation and generation.



[CoRL 2024] "RAM: Retrieval-Based Affordance Transfer for Generalizable Zero-Shot Robotic Manipulation." Kuang et al.

[In submission 2025] "Learning from Massive Human Videos for Universal Humanoid Pose Control." Mao et al.

Robot learning from non-robot data.

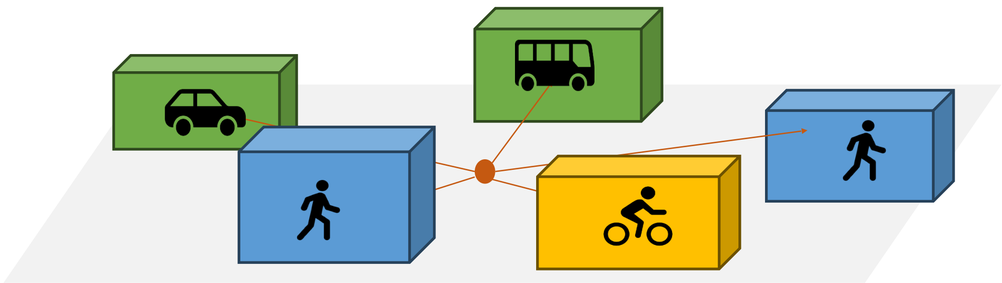
From understanding to editing

Can we speed up a car?

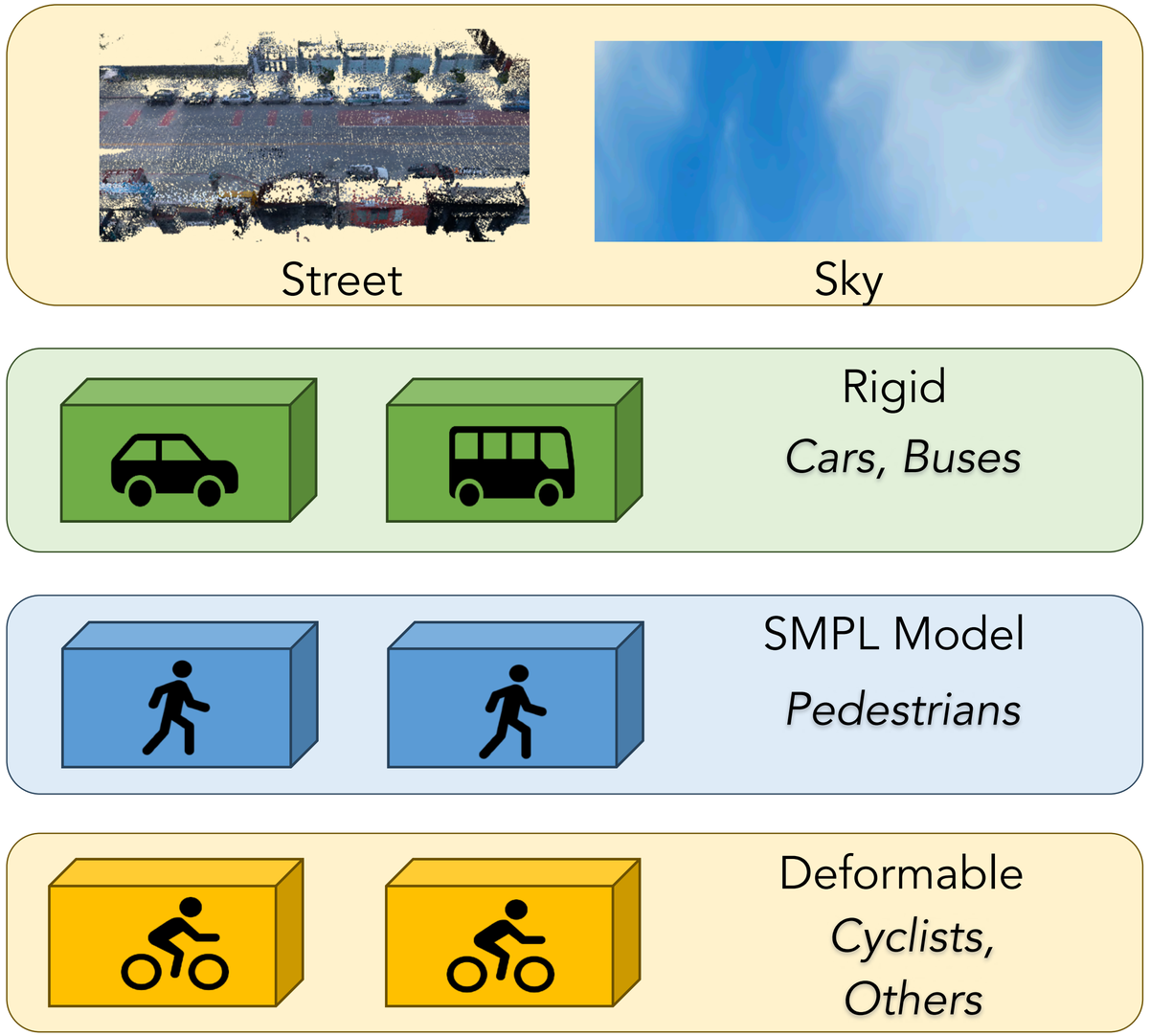
Can we remove a pedestrian from the scene?

[ICLR 2025] “Omni Urban Scene Reconstruction.” Chen et al.

Scene Modeling



Gaussian Scene Graph



Applications Let People Dance!

Applications Driving Simulation

Applications Bullet time

1. Semantic and physical understanding.
2. 3D reconstruction and simulation.
3. **Learning from non-robotic data**.

- Robotic data (real-world or simulation)
- Hand-object interaction (HOI) data
- Internet-scale videos
- AIGC (DALL·E, Sora…)
- Even cartoon or sketches!

- Contain rich actionable knowledge.
- But the data themselves are noisy and heterogeneous.

[In submission 2025] “Robot Learning from Any Images” Zhao et al.

Robot learning from non-robotic data

Manipulate **unseen objects** in **unseen environments** with **unseen embodiments.**

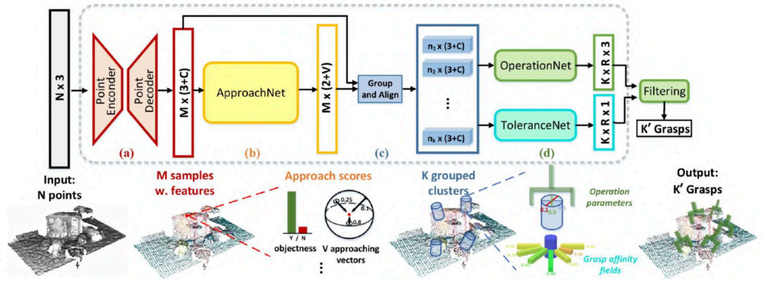
Learn manipulation from costly **in-domain demonstrations**.

Acquire versatile manipulation capabilities from abundant **out-of-domain data**.

...

RAM: Retrieval-Based Affordance Transfer for Generalizable Zero-Shot Robotic Manipulation

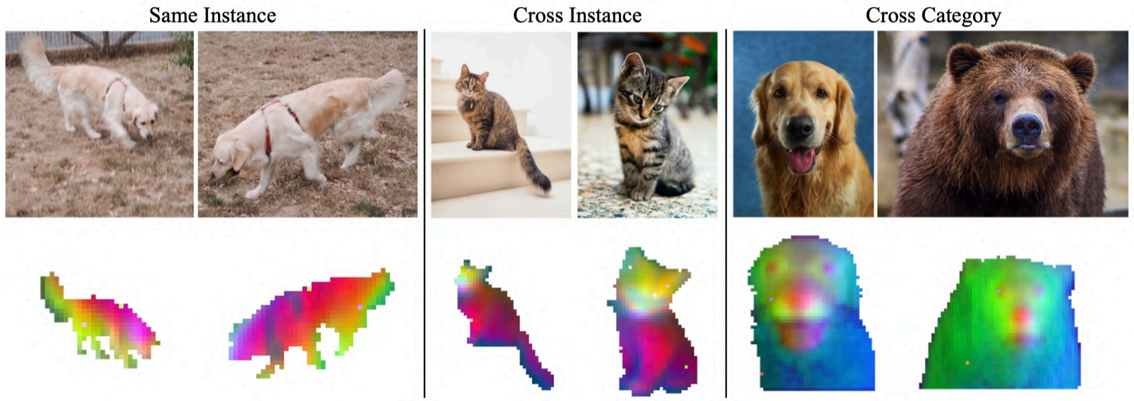
- Represent the actionable knowledge as **transferrable affordance**, i.e. ‘where’ and ‘how’ to act
  - ‘where’ to act: 3D contact point
  - ‘how’ to act: 3D post-contact direction
- Off-the-shelf grasp generators and motion planners for execution.



[CoRL 2024] “RAM: Retrieval-Based Affordance Transfer for Generalizable Zero-Shot Robotic Manipulation.” Kuang et al.

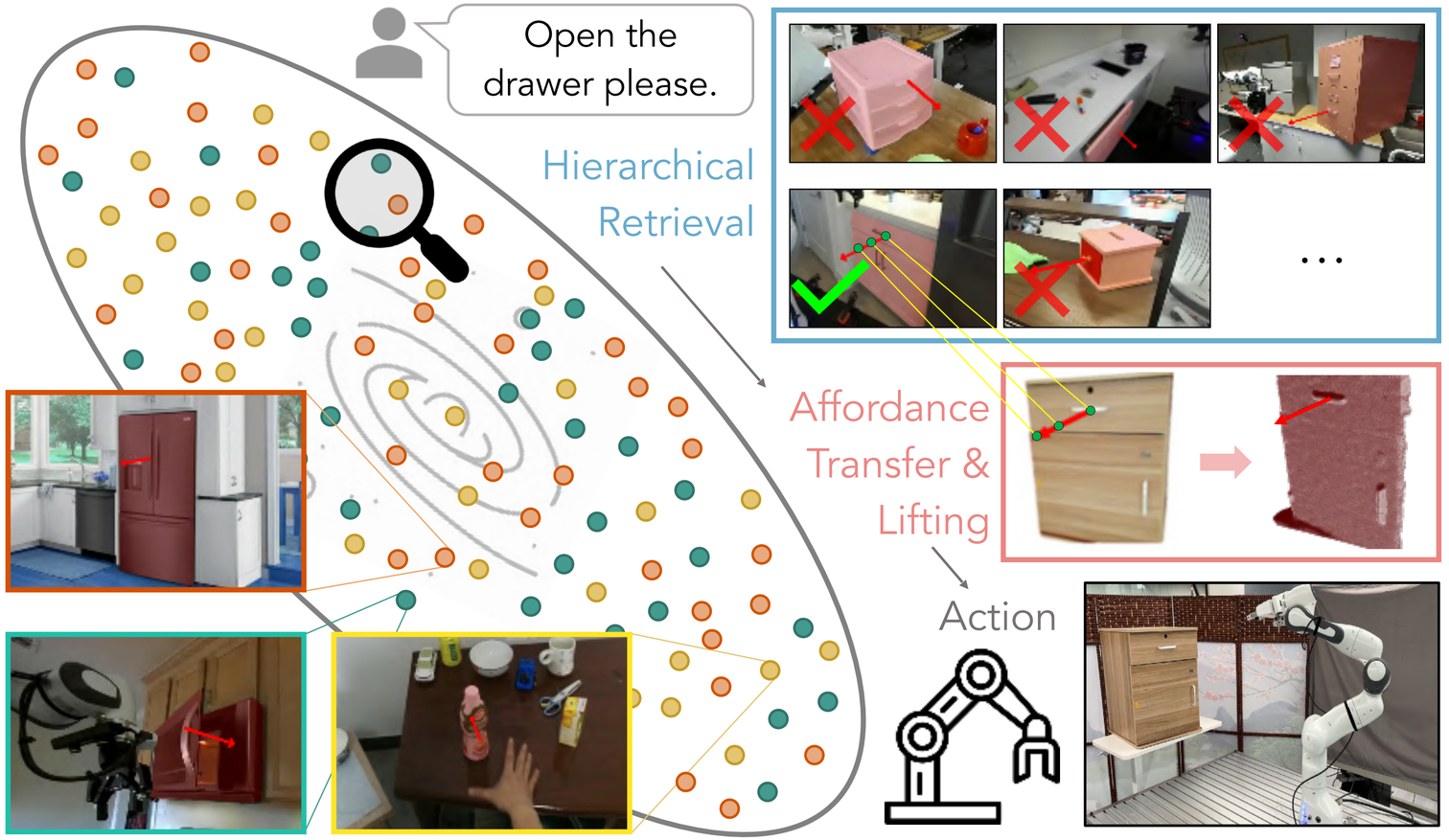
How to match points with the same semantic concepts?

- Emergent dense correspondence of feature maps.
- Cross domain/instance/category generalization.



[NeurIPS 2023] “Emergent correspondence from image diffusion.” Tang et al.

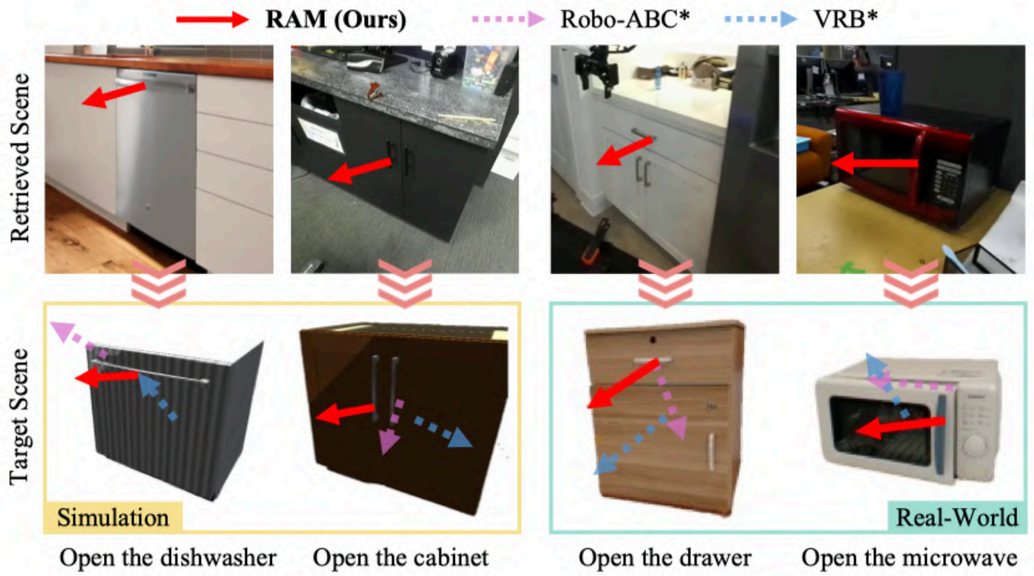
Overview



The Pipeline of RAM.

Experiment Results

Object														AVG
Task	O	C	O	C	O	C	O	O	O	P	P	P	P	/
Where2Act [14]	2	34	2	54	2	**68**	2	0	/	/	/	/	/	20.50
VRB* [12]	8	62	6	56	16	66	4	12	10	18	28	44	60	30.77
Robo-ABC* [44]	20	58	22	60	30	46	30	28	26	40	54	66	60	41.54
RAM (Ours)	**38**	**68**	**32**	**76**	**32**	50	**66**	**54**	**38**	**46**	**56**	**72**	**64**	**52.62**



Object							AVG
Task	O	O	O	P	P	P	/
Robo-ABC* [44]	2/5	1/5	1/5	**3/5**	**4/5**	4/5	50.0
RAM (Ours)	**3/5**	**2/5**	**3/5**	**3/5**	**4/5**	**5/5**	**66.7**

Zero-Shot Manipulation
"Open the microwave"

Retrieved Demonstration

3x

Zero-Shot Manipulation
"Open the cabinet"

Retrieved Demonstration

3x

Zero-Shot Manipulation
"Pick up the banana"

Retrieved Demonstration

3x

One-Shot Visual Imitation
"Pick up the tissue paper"

3x

One-Shot Visual Imitation
"Pick up the tissue box"

3x

One-Shot Visual Imitation
"Close the drawer"

3x

More DoFs
Not easily handled by affordance
Action retargeting is hard

How can we learn humanoid dexterity from Internet data?

UH-1: Learning from Massive Human Videos for Universal Humanoid Pose Control

Massive Internet Videos

YouTube

Youtube

DeepMind

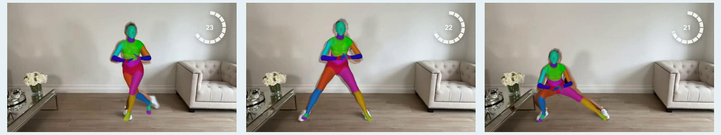
Kinetics 700

Allen Institute for AI

Charades

Video Clip Extraction

3D Human Pose Estimation



Video Captioning

"A young woman is doing a workout in a living room, using her legs and arms to perform various exercises."

Goal-based Reinforcement Learning

Motion Retargeting from Humans to Humanoids

Text Representation

"A young woman is doing a workout in a living room, using her legs and arms to perform various exercises."

Humanoid Robot Actions

UH-1 Transformer

Humanoid Action Tokens

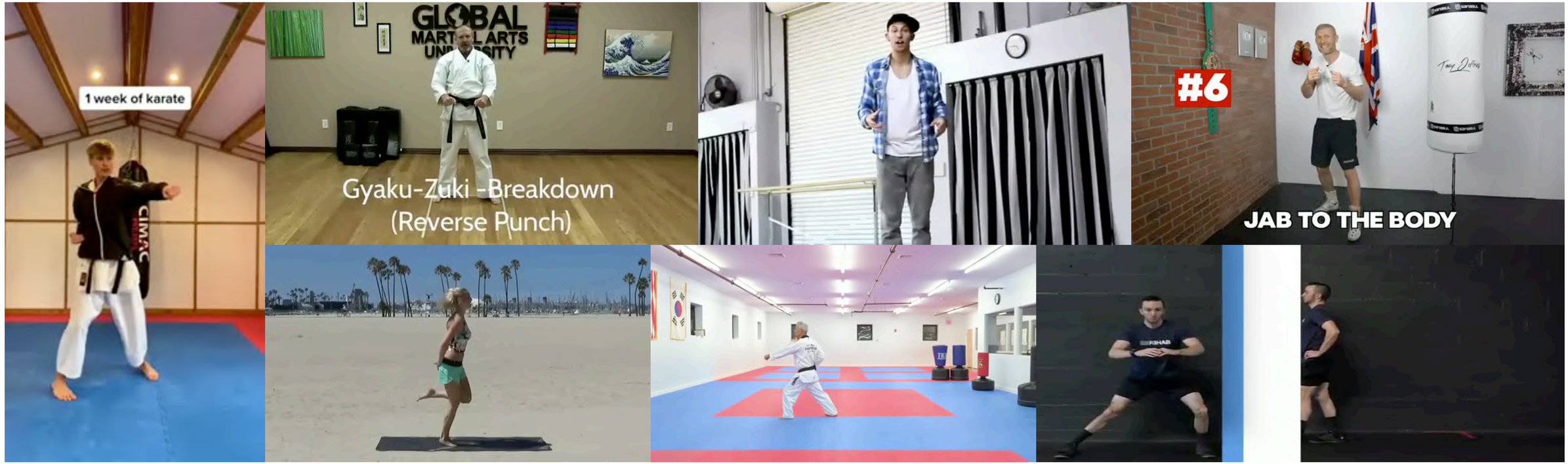
An automatic humanoid data engine

A unified whole-body control model

[In submission 2025] "Learning from Massive Human Videos for Universal Humanoid Pose Control." Mao et al.

Data Collection

We collect 163, 800 video clips from diverse sources.



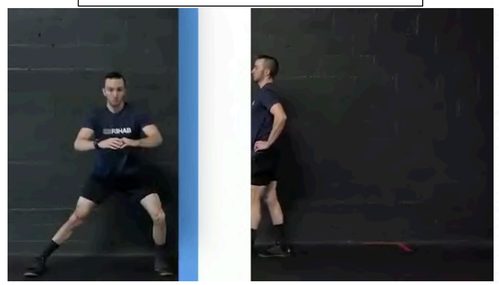
Data Collection

Videos are further annotated with captioning tools.

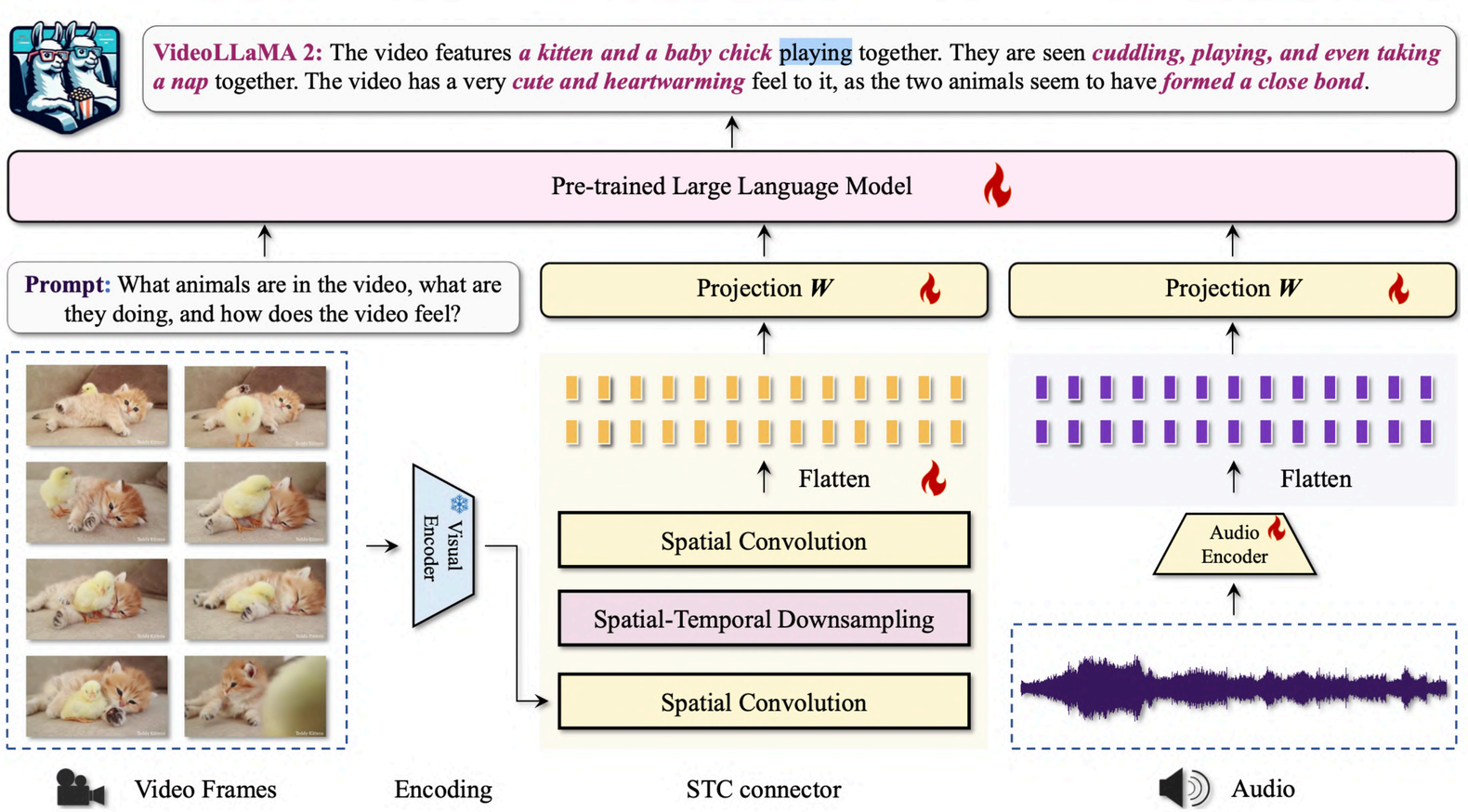
"practicing martial arts, standing."

"standing and speaking."

"practicing yoga."

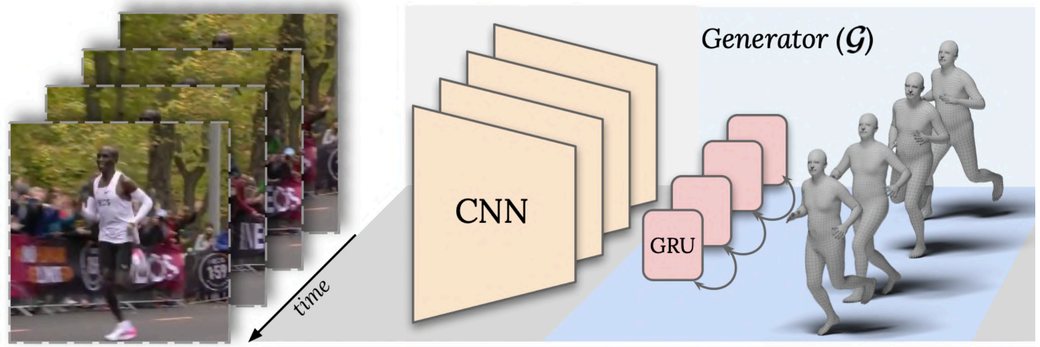


"doing squats, lunges, and jumping jacks."



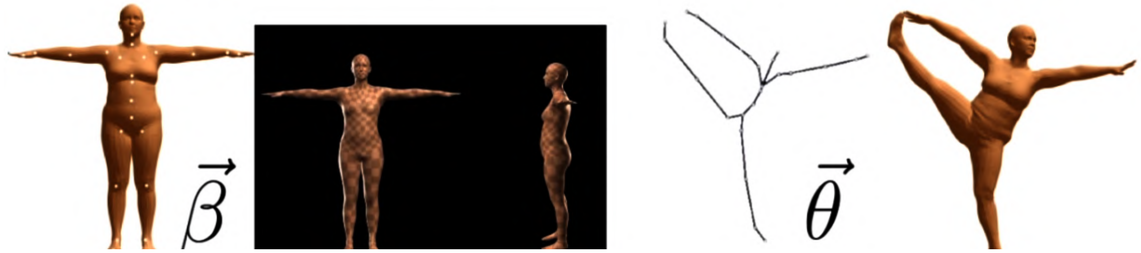
[Preprint 2024] "VideoLLaMA 2: Advancing Spatial-Temporal Modeling and Audio Understanding in Video-LLMs." Cheng et al.

Human Motion Representation

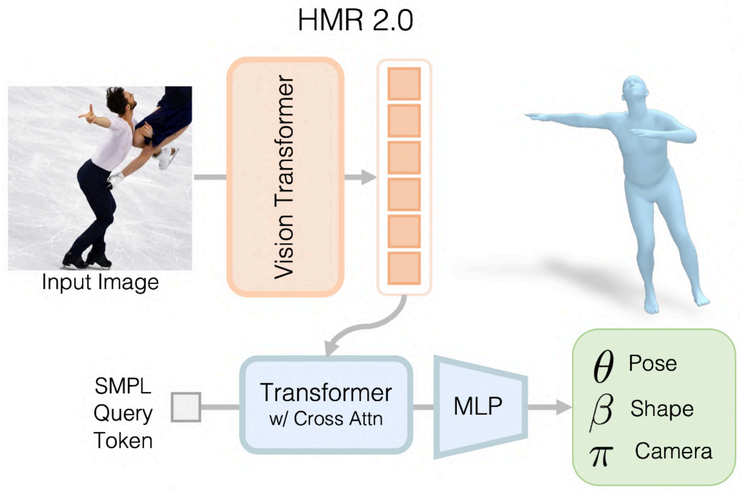


Shape: PCA coefficients

Pose: Rotation of 23 joints

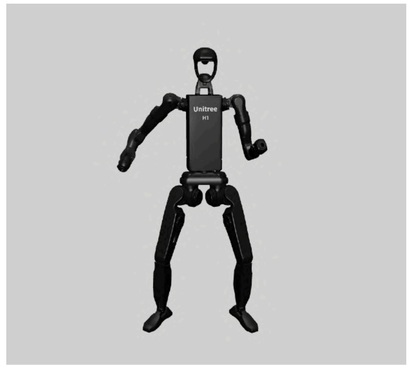


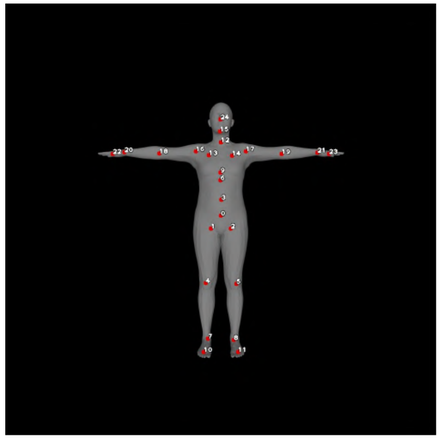
SMPL Model

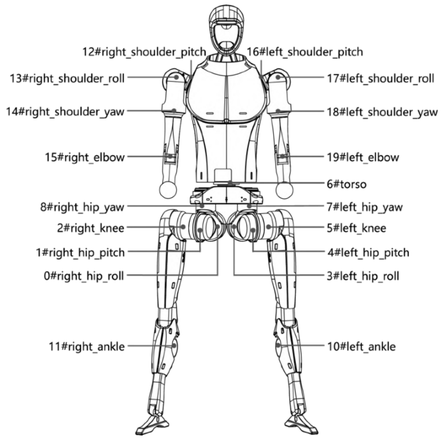


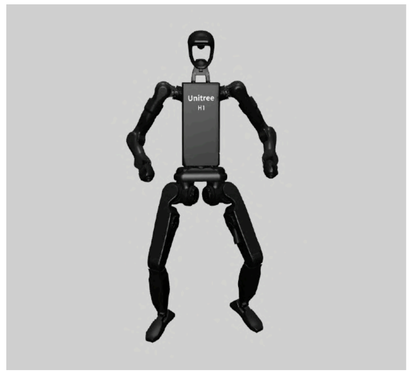
[Kocabas et al., CVPR 2020]
[Goel et al., CVPR 2024]

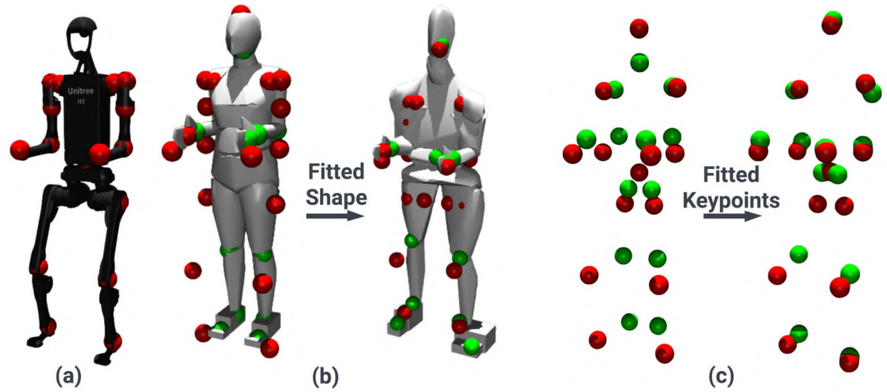
Human-to-Humanoid Motion Retargeting







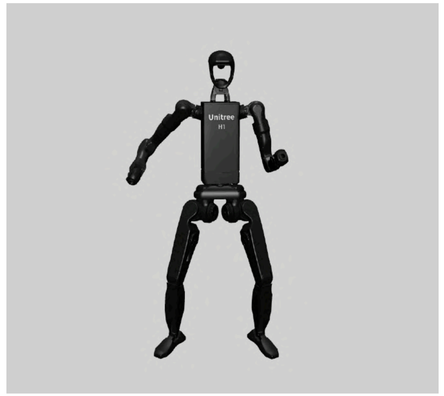


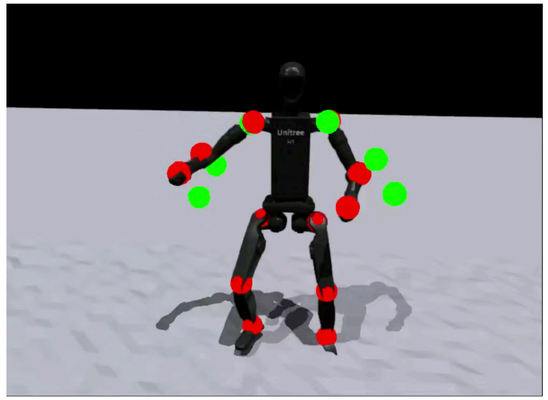


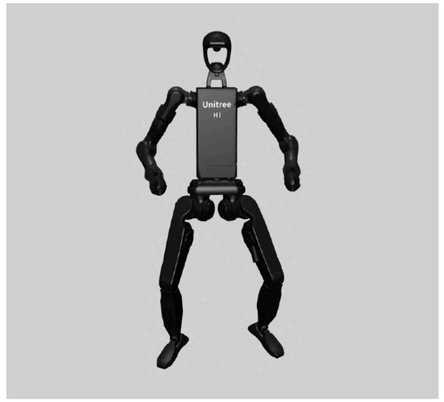
$$\min_{\beta} ||\mathcal{P}^T_{joints} - \mathcal{P}^T_{robot}||_2,$$

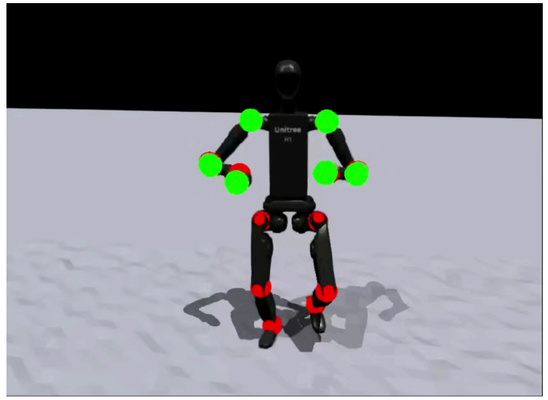
$$\text{s.t.} \quad \mathcal{P}^T_{joints} = F_{fk}(\mathcal{P}_{human}(\beta, \theta^T, t_{root})),$$

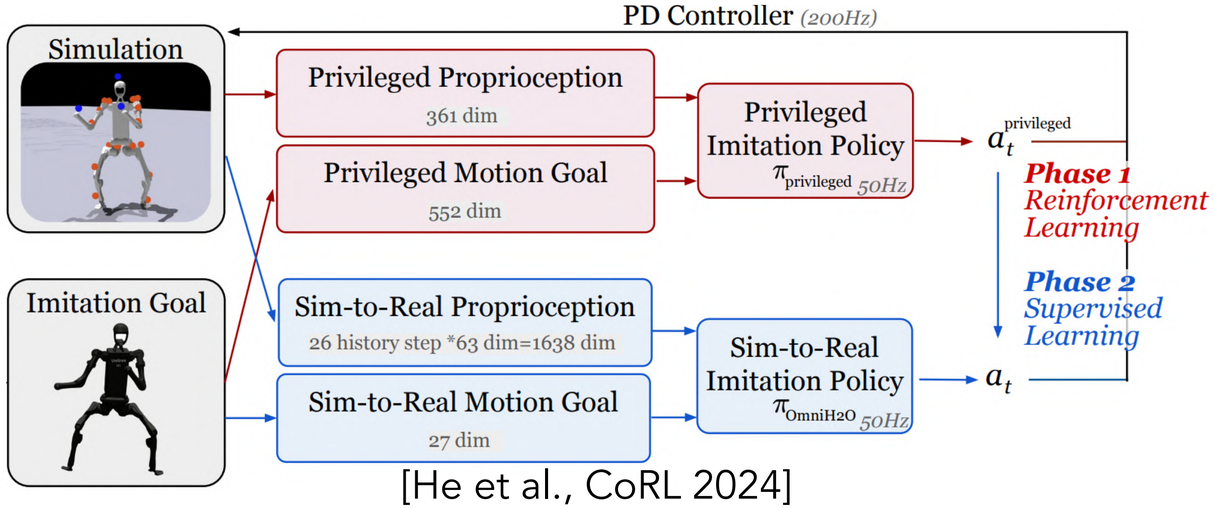
Sim-to-Real Adaptation





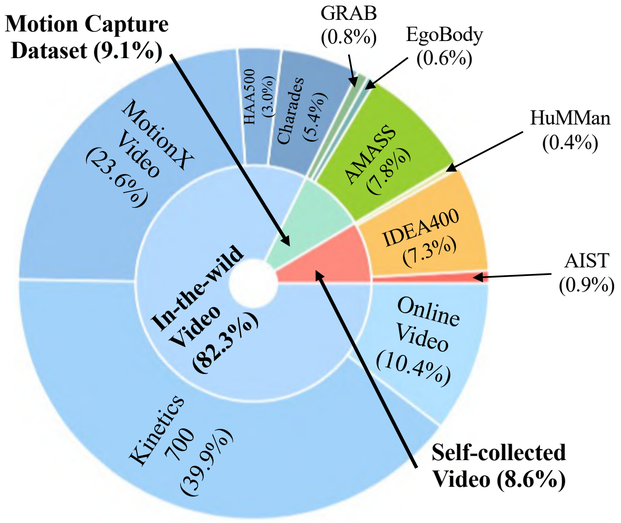




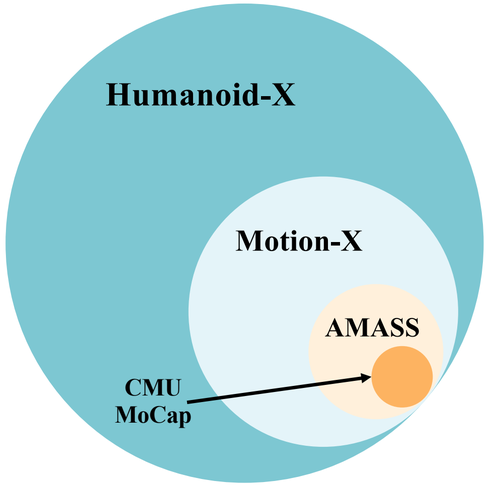


[He et al., CoRL 2024]

Dataset

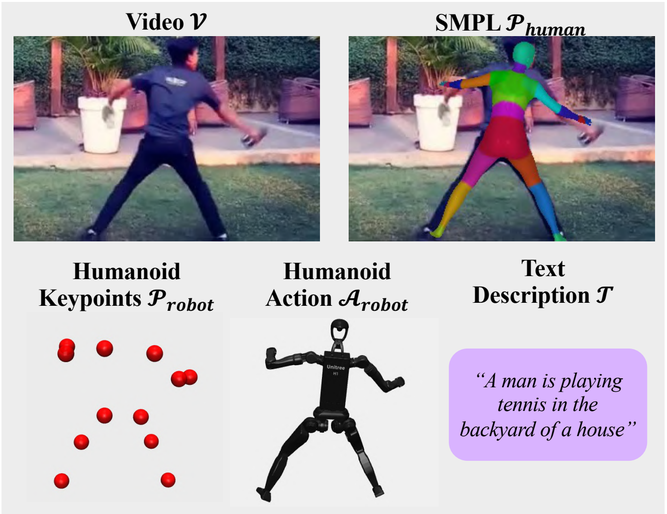


(a) Data Distribution



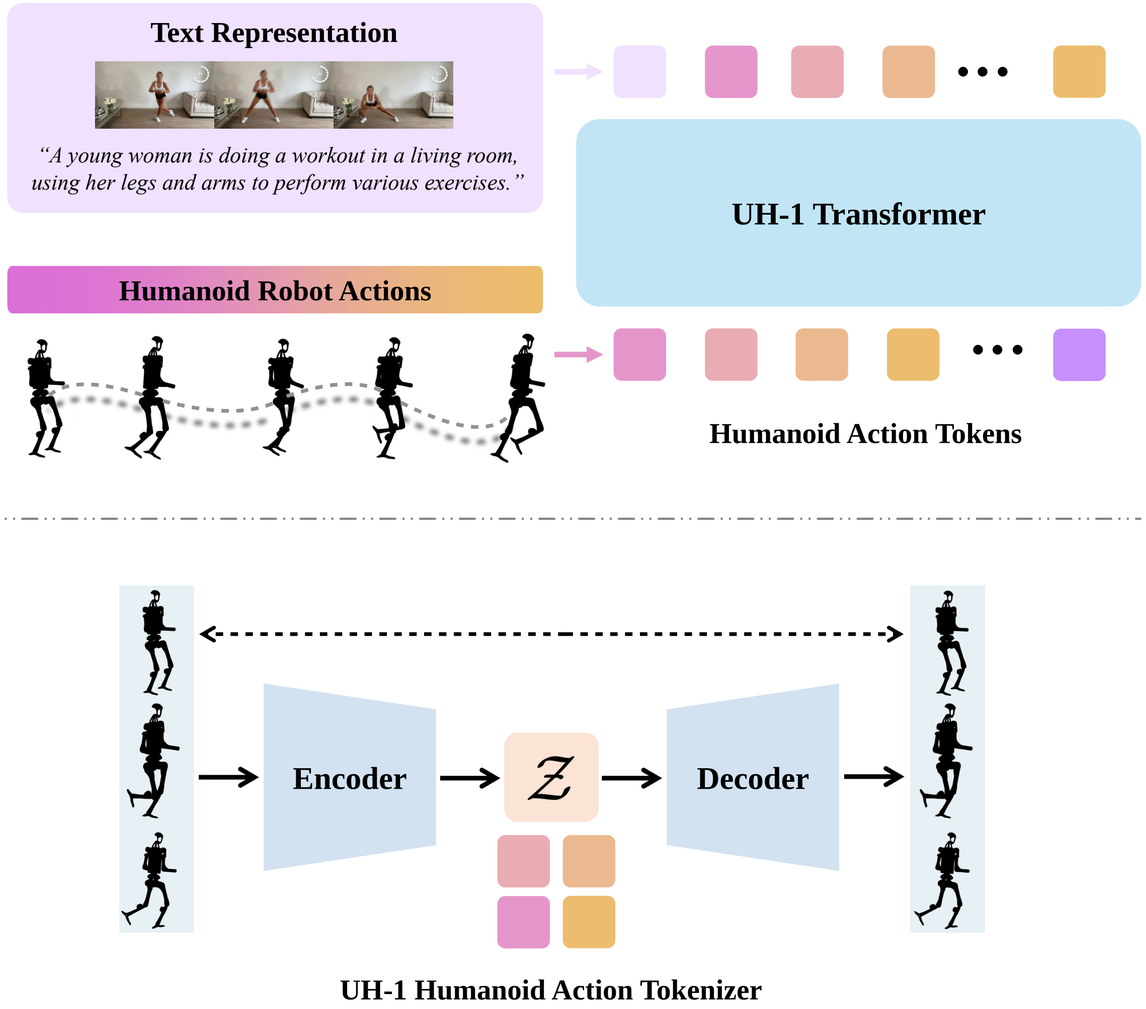
(b) Data Scale

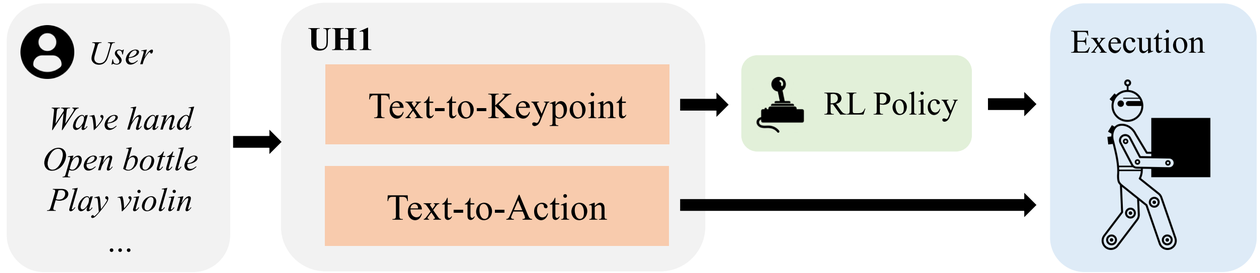
(c) Vocabulary Diversity



(d) Motion Sample

Universal Humanoid (UH-1) Architecture





One Model, Two Control Modes

Research Questions

- **Universal Pose Control with UH-1**: Does UH-1 model enable universal humanoid robot pose control based on text commands?

- **Scalability and Generalization with Humanoid-X**: Does the large-scale Humanoid-X dataset facilitate scalable training and improve the generalization ability of UH-1?

- **Real-World Deployment of UH-1**: Can UH-1 model be deployed on real humanoid robots to enable reliable robotic control in real-world environments?

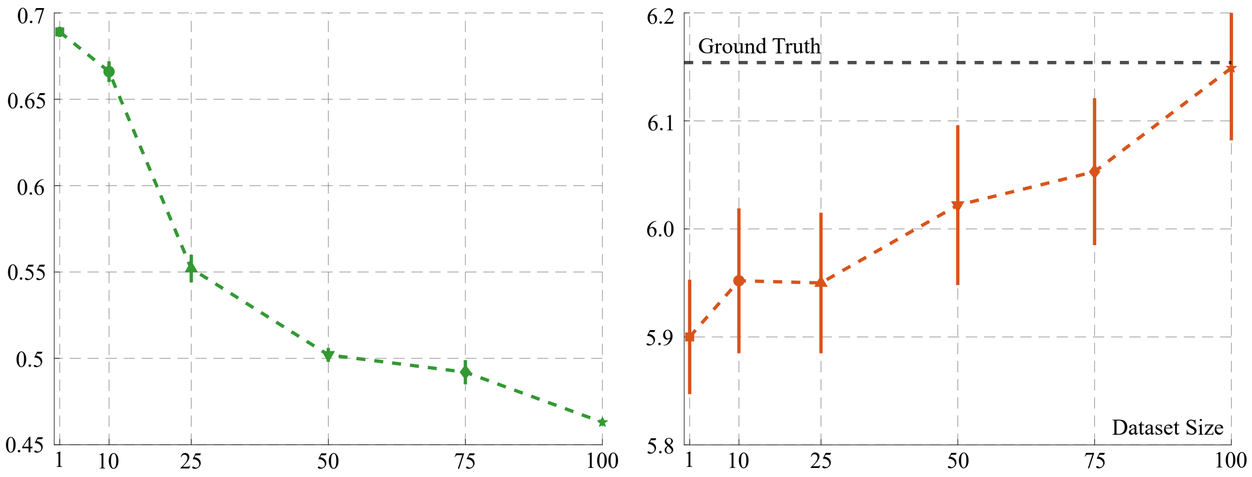
Universal Pose Control with UH-1

- Baseline models: Motion Diffusion Model (MDM) and Text-to-Motion GPT (T2M-GPT)

Methods	**FID ↓**	**MM Dist ↓**	**Diversity ↑**	**R Precision ↑**
Oracle	$0.005^{\pm.001}$	$3.140^{\pm.010}$	$9.846^{\pm.062}$	$0.780^{\pm.003}$
MDM [57]	$0.582^{\pm.051}$	$5.921^{\pm.034}$	$10.122^{\pm.078}$	$0.617^{\pm.007}$
T2M-GPT [71]	$0.667^{\pm.109}$	$3.401^{\pm.017}$	$\mathbf{10.328^{\pm.099}}$	$0.734^{\pm.004}$
UH-1 (ours)	$\mathbf{0.445^{\pm.078}}$	$\mathbf{3.249^{\pm.016}}$	$10.157^{\pm.106}$	$\mathbf{0.761^{\pm.003}}$

Scalable Learning with Humanoid-X

- Increasing data size leads to consistent performance improvement.
- Pre-training on Humanoid-X helps generalization.

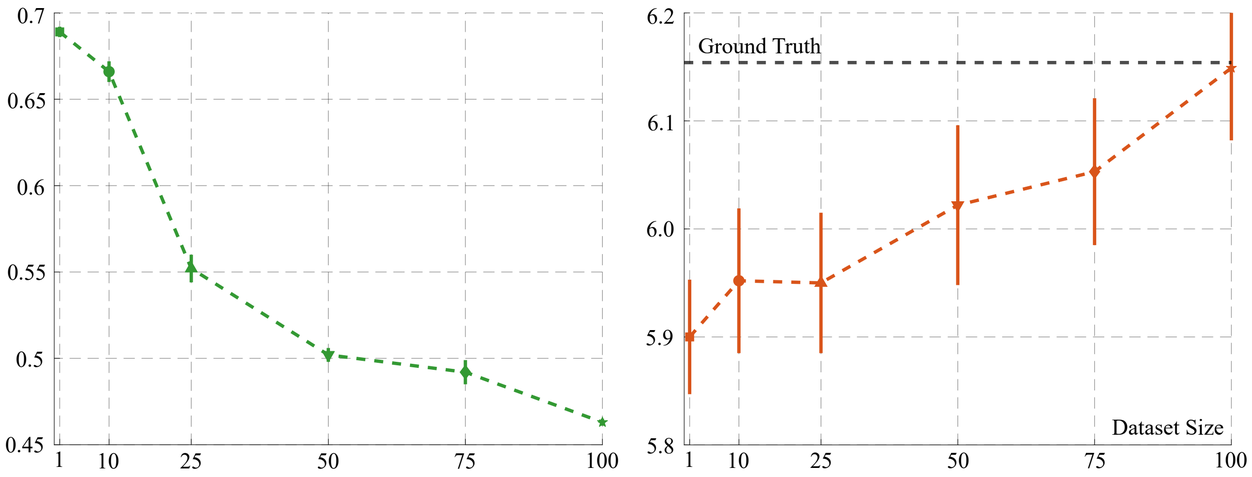


(a) FID ↓ (b) Diversity ↑

Dataset	FID ↓	MM Dist ↓	Diversity ↑	R Precision ↑
Oracle	$0.005^{\pm.001}$	$3.140^{\pm.010}$	$9.846^{\pm.062}$	$0.780^{\pm.003}$
HumanoidML3D	$0.445^{\pm.078}$	$3.249^{\pm.016}$	$10.157^{\pm.106}$	$0.760^{\pm.003}$
Humanoid-X	$\mathbf{0.379}^{\pm.046}$	$\mathbf{3.232}^{\pm.008}$	$\mathbf{10.221}^{\pm.100}$	$\mathbf{0.761}^{\pm.003}$

Scalable Learning with Humanoid-X

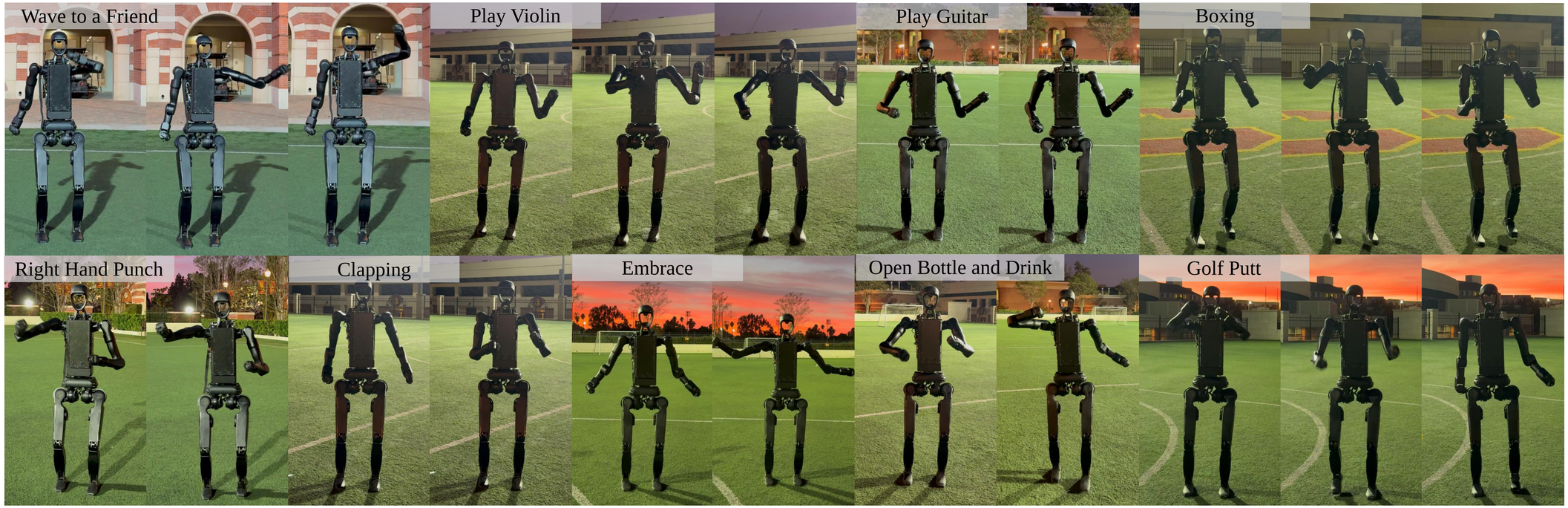
- Increasing data size leads to consistent performance improvement.
- Pre-training on Humanoid-X helps generalization.



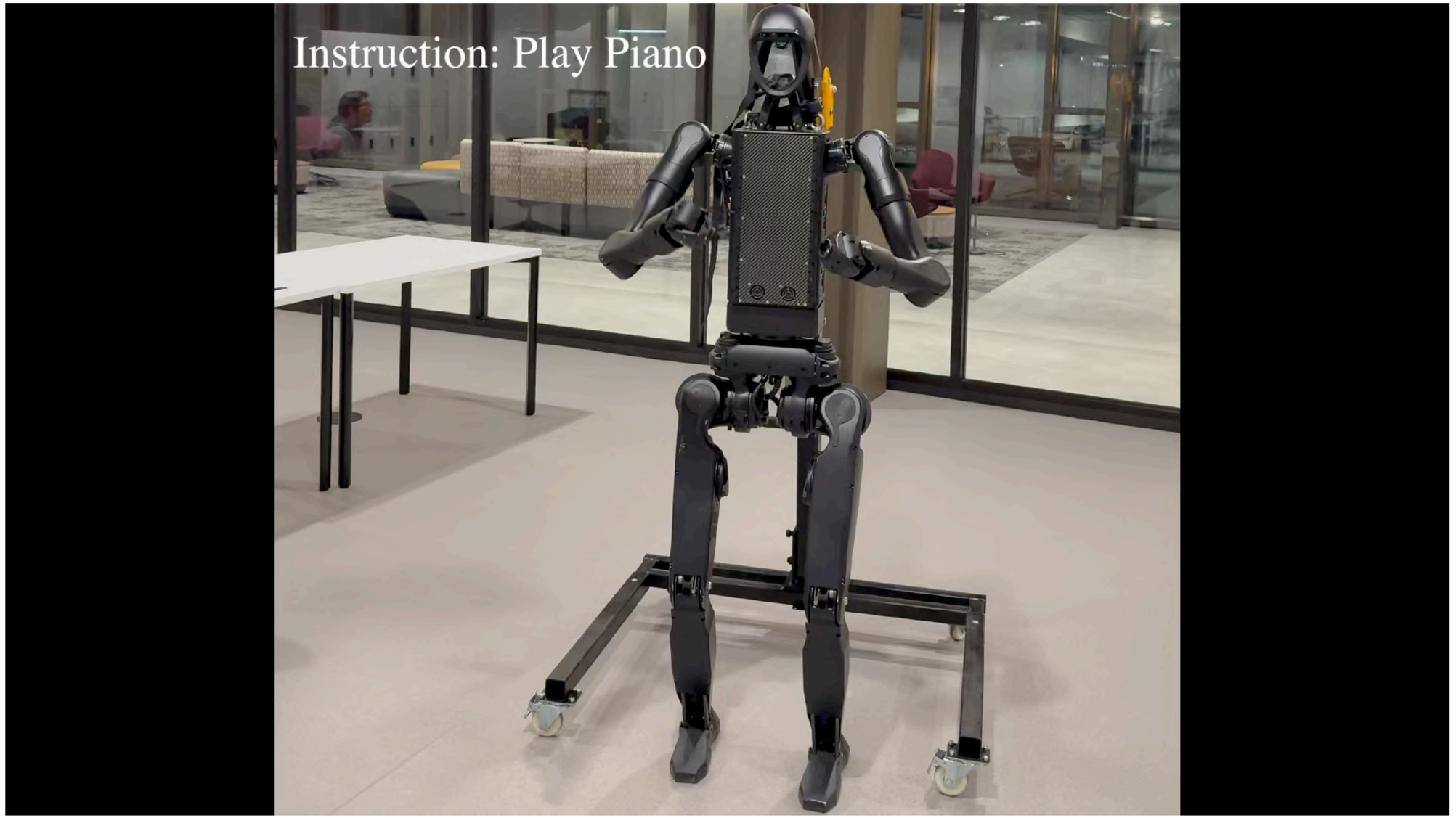
(a) FID ↓ (b) Diversity ↑

Dataset	FID ↓	MM Dist ↓	Diversity ↑	R Precision ↑
Oracle	$0.005^{\pm.001}$	$3.140^{\pm.010}$	$9.846^{\pm.062}$	$0.780^{\pm.003}$
HumanoidML3D	$0.445^{\pm.078}$	$3.249^{\pm.016}$	$10.157^{\pm.106}$	$0.760^{\pm.003}$
Humanoid-X	$\mathbf{0.379}^{\pm.046}$	$\mathbf{3.232}^{\pm.008}$	$\mathbf{10.221}^{\pm.100}$	$\mathbf{0.761}^{\pm.003}$

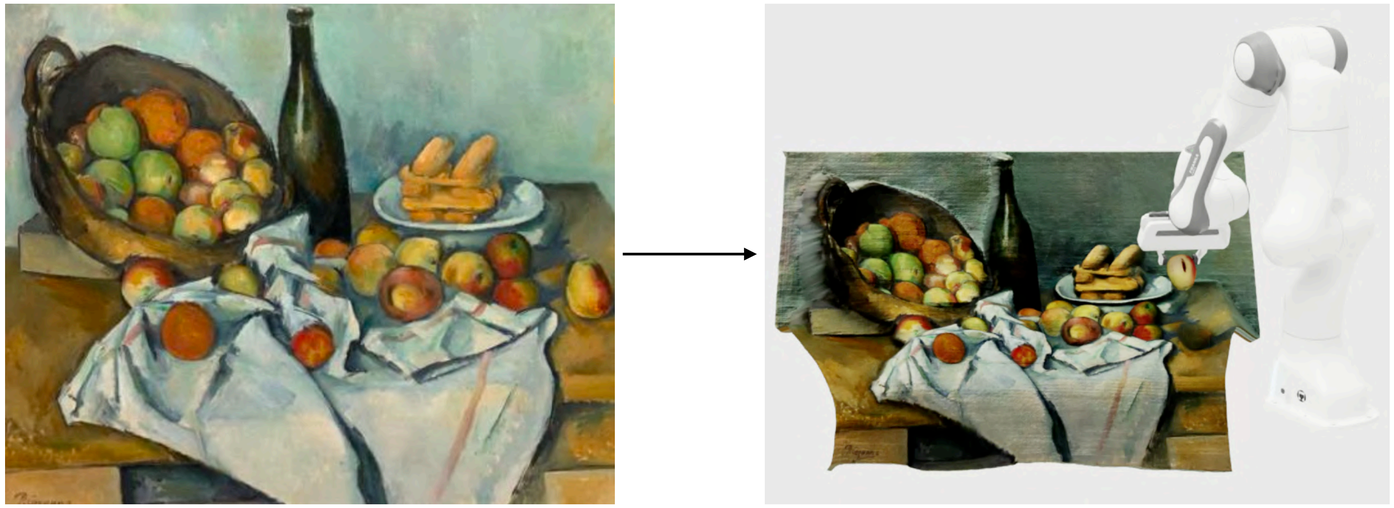
Real-World Deployment of UH-1



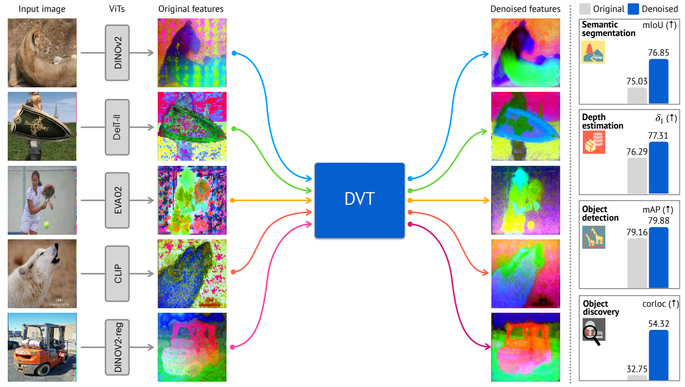
Real-World Deployment of UH-1

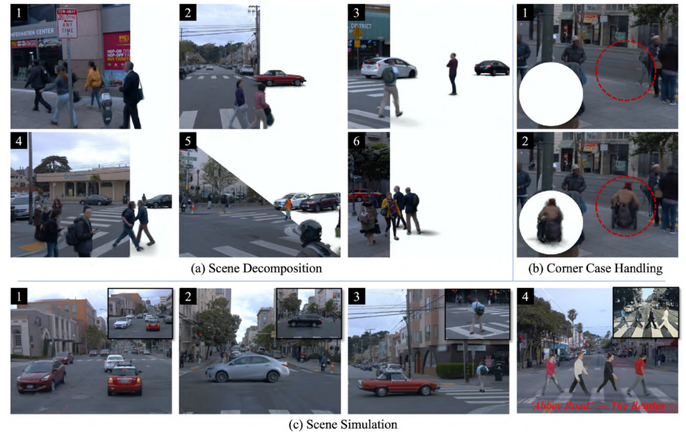


Summary: from vision to action with minimal supervision.

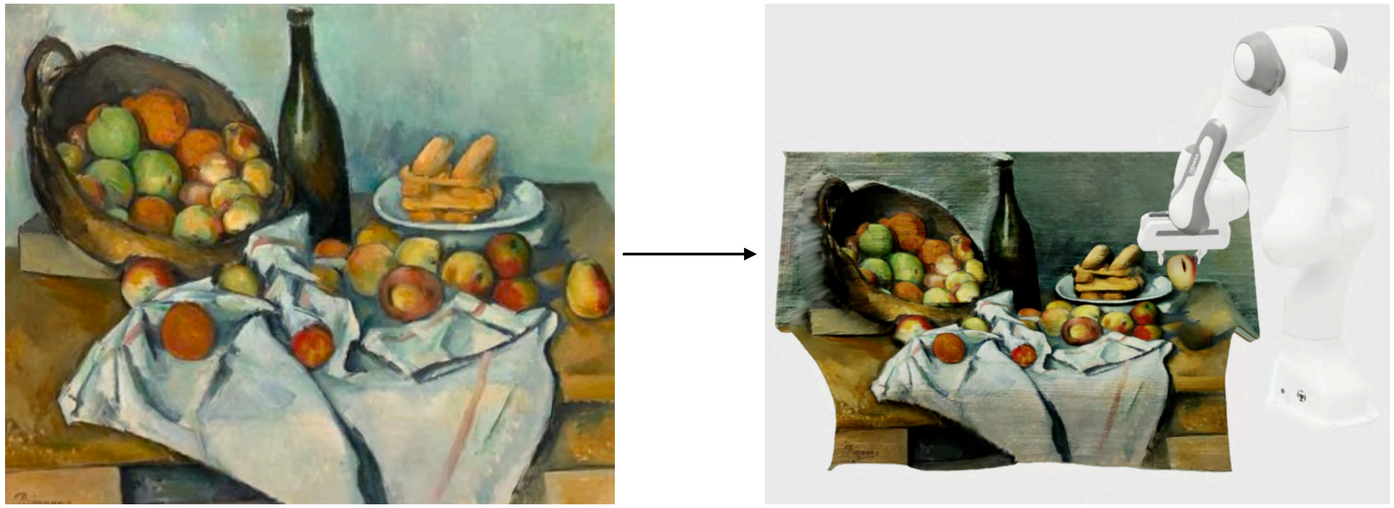


1. Semantic and physical understanding.
2. 3D reconstruction and simulation.
3. Learning from non-robotic data.

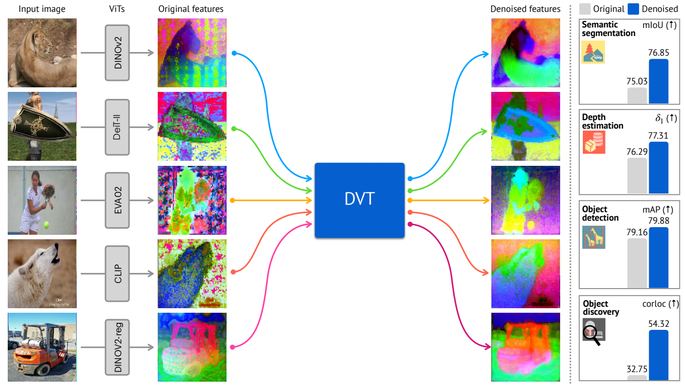


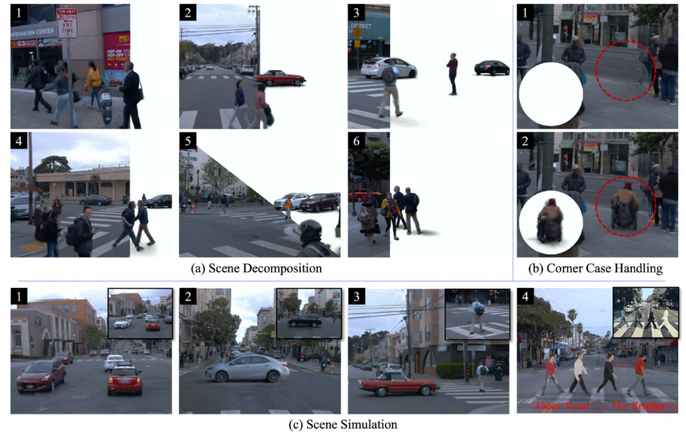


Summary: from vision to action with minimal supervision.

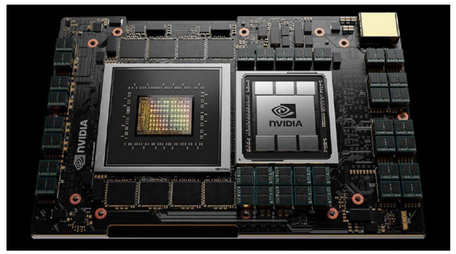


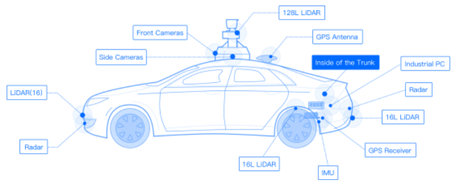
1. Data is still scarce.
2. Robotic algorithms are not end-to-end.
3. Hardware design is more control focused.





Goal: enable embodied intelligence to solve hard problems





• • •

Hardware

ROS

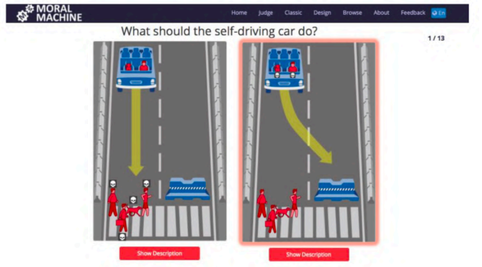
PyTorch TensorFlow

Perception Planning Control

• • •

Software

Privacy

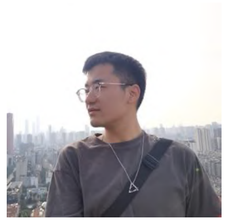


Ethics

• • •

Society

Acknowledgement


Junjie Ye


Jiageng Mao


Jiawei Yang


Siheng Zhao

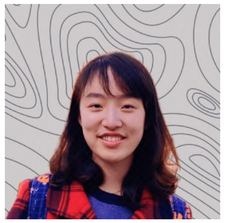

Cameron Smith


Yuxuan Kuang


Wei Zhou


Boris Ivanovic


Sanja Fidler


Congyue Deng


Zan Gojcic


Marco Pavone


Vitor Guizilini


Leo Guibas