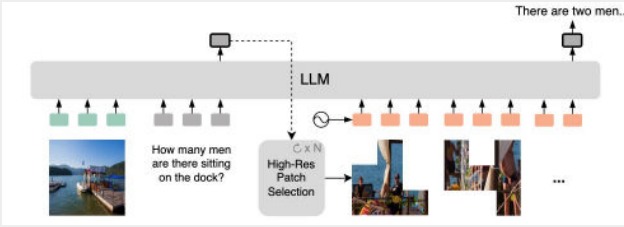
The Future of Visual AI: Efficient Multimodal Intelligence

or

(Efficient) LLMs from Text to Vision and Robotics and back…

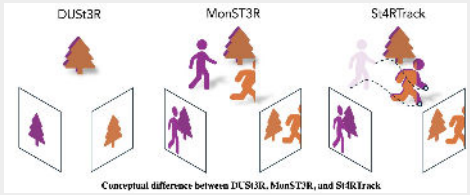
Prof. Trevor Darrell

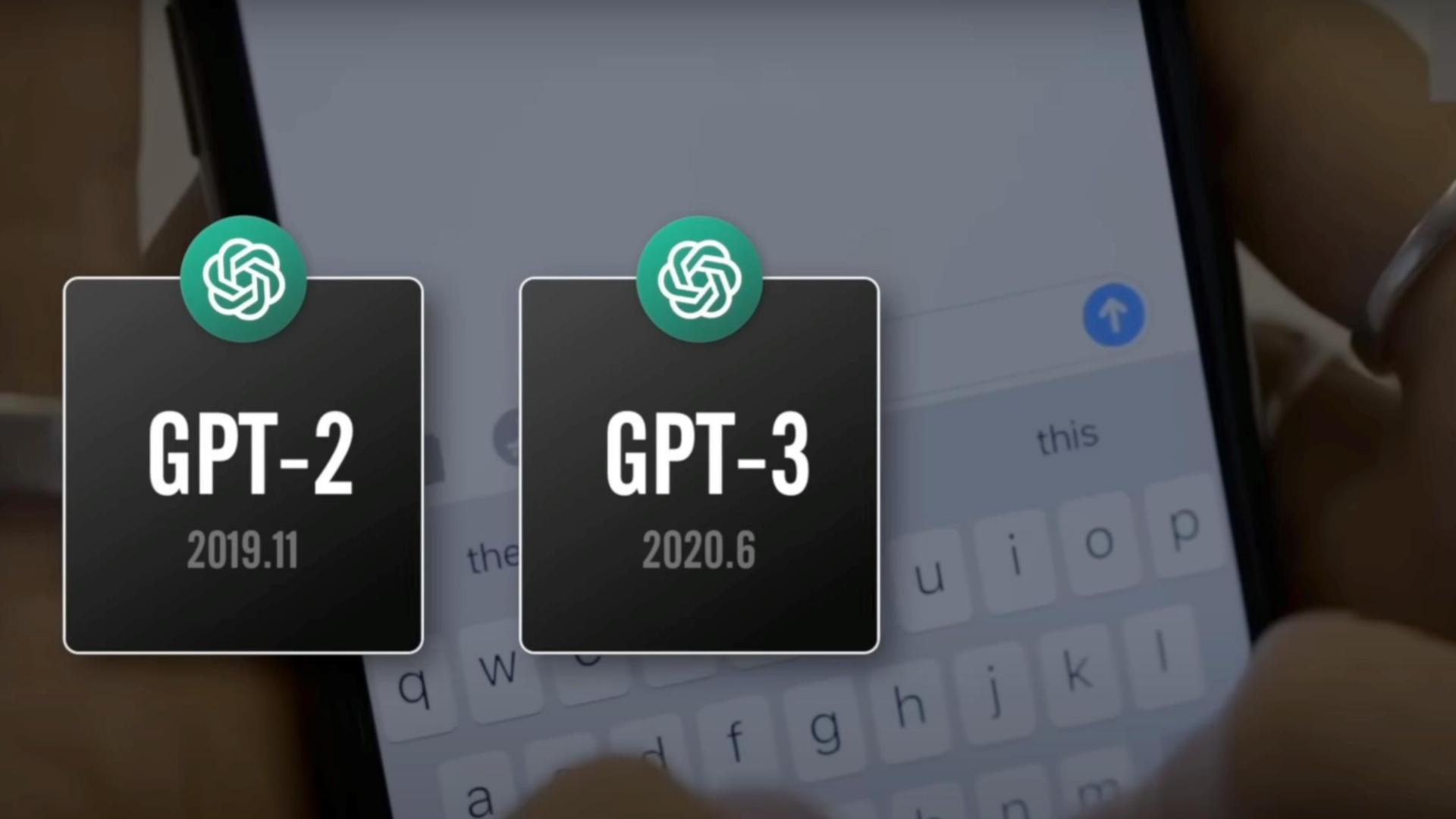
UC Berkeley



$$J(s_{\tau:\tau+m}, s^*, a_{\tau:\tau+m-1}) = \text{Distance to Goal} + \text{Action Constraints} + \text{State Constraints}$$

States Goal Actions




GPT-2
2019.11
GPT-3
2020.6
this
the

LLMs for … ?

Text

Images/Videos

Actions

Sounds

…

LLMs from Text to Vision and Robotics and back…

- **Are LLMs Grounded?**
- Reducing VLM Hallucination
- Efficient Scaling of VLMs
- Visual Tokens for Non-linguistic Generation
- Navigation World Models
- 4D Reconstruction for Humanoid Robotics

A vision persons view of LLMs (circa 2022)

+ Amazing at textual translation and reasoning!
+ An interesting space to define visual tasks?

- Definitely not “grounded” in any *real* way, right?

Various “Chinese Room” and “Stochastic Parrot” arguments ensue…

Prompt: A gray cat and an orange dog on the grass

— Stable Diffusion

Where is the orange dog?

orange dog ≠ orange

Prompt: A blue cube directly above a red cube with a vase on the left of them

— Stable Diffusion

No red cube or vase in the designated location?

Prompt: A wooden table without bananas

— Stable Diffusion

"Without" bananas, *not* "with"?

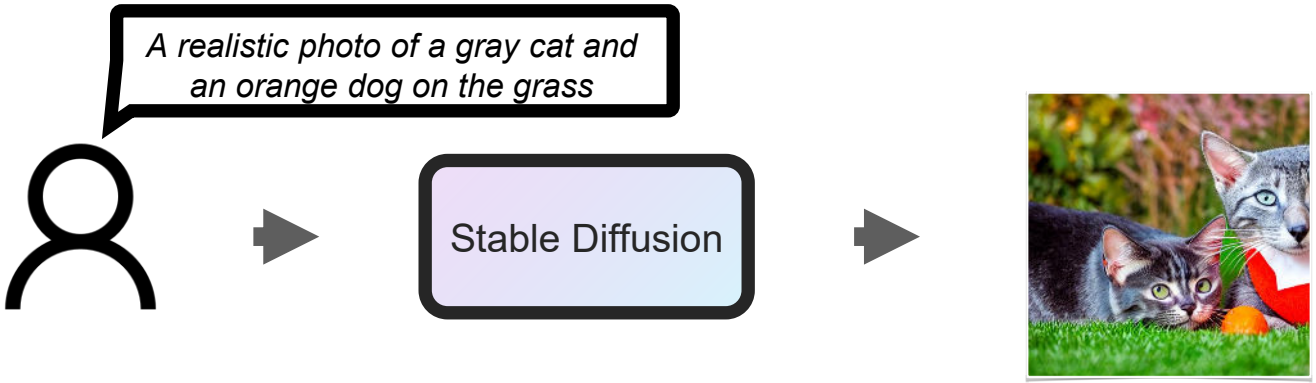
Prompt: A man in red standing next to another woman in blue

— Stable Diffusion

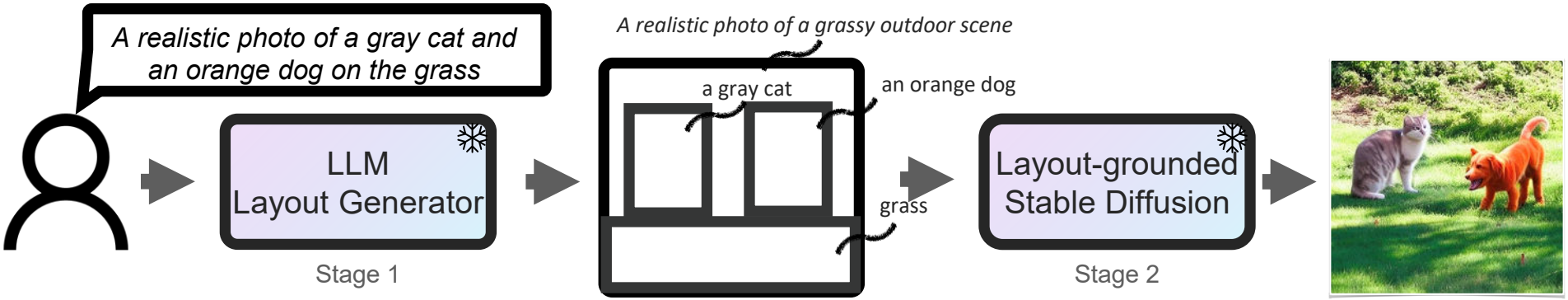
Is the man in red?

Is the woman in blue?

Typical Text-to-Image Diffusion



LLM-grounded Diffusion (LMD)



LMD is *training-free*

arXiv > cs > arXiv:2305.13655

Search...

Help | Ad

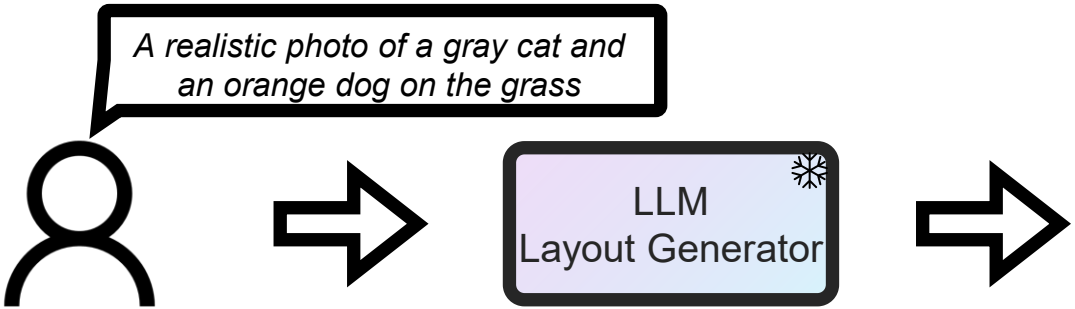
Computer Science > Computer Vision and Pattern Recognition

[Submitted on 23 May 2023 (v1), last revised 4 Mar 2024 (this version, v3)]

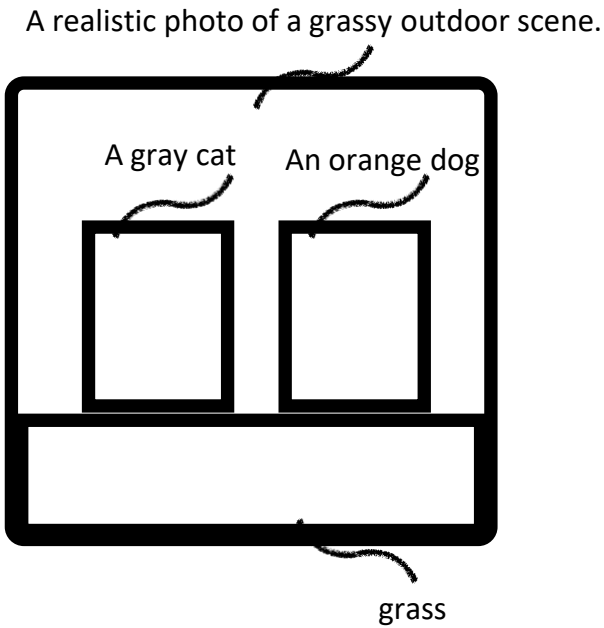
LLM-grounded Diffusion: Enhancing Prompt Understanding of Text-to-Image Diffusion Models with Large Language Models

Long Lian, Boyi Li, Adam Yala, Trevor Darrell

Layout



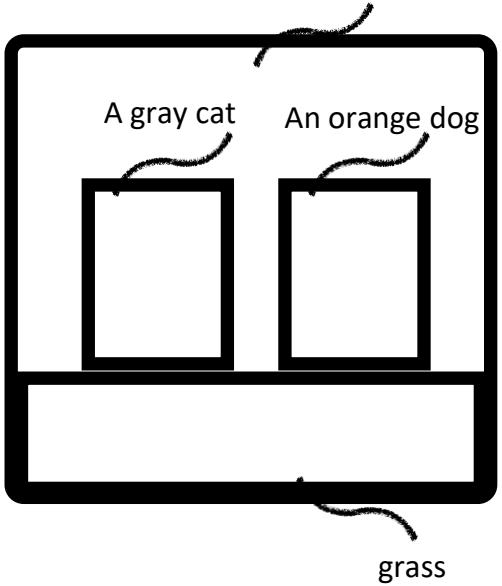
Stage 1



Your task is to generate the bounding boxes for the objects mentioned in the caption, along with a background prompt describing the scene…
[In-context examples]
Caption: A realistic photo of a gray cat and an orange dog on the grass
Objects: [('a gray cat', [50, 120, 180, 200]), ('an orange dog', [300, 120, 180, 200]), ('grass', [0, 340, 512, 172])]
Background prompt: A realistic photo of a grassy outdoor scene

⇨ Image

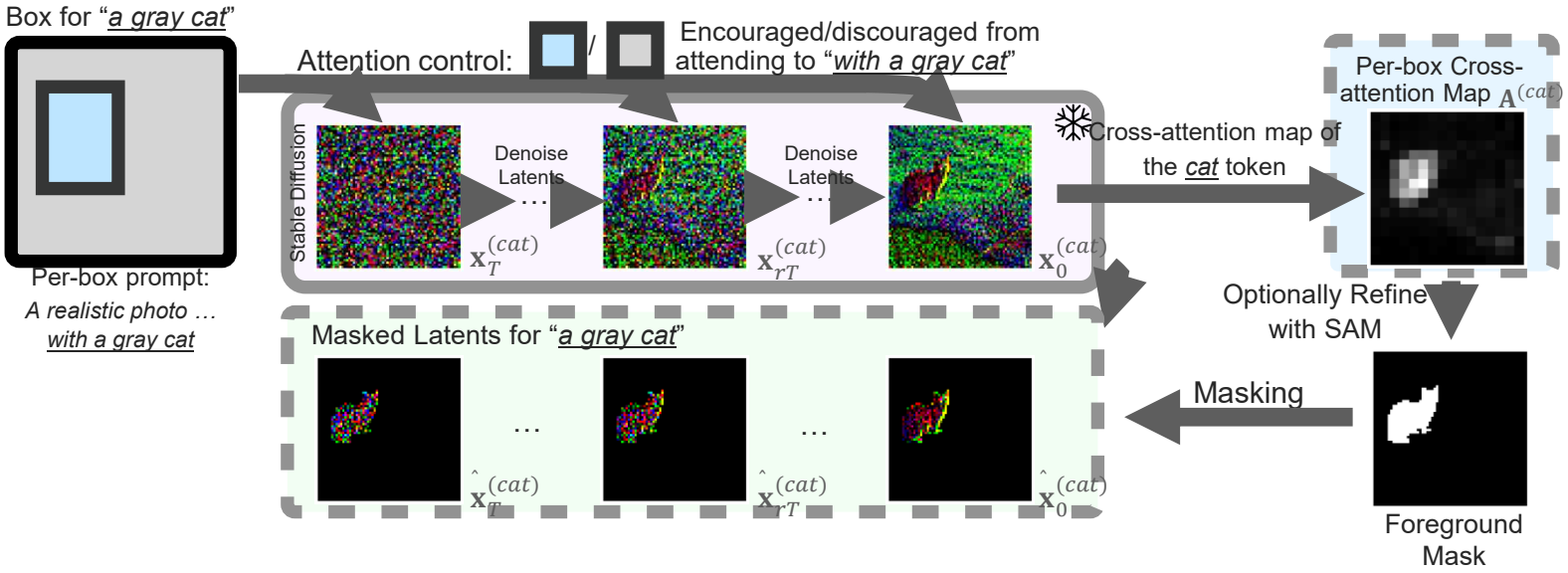
A realistic photo of a grassy outdoor scene.



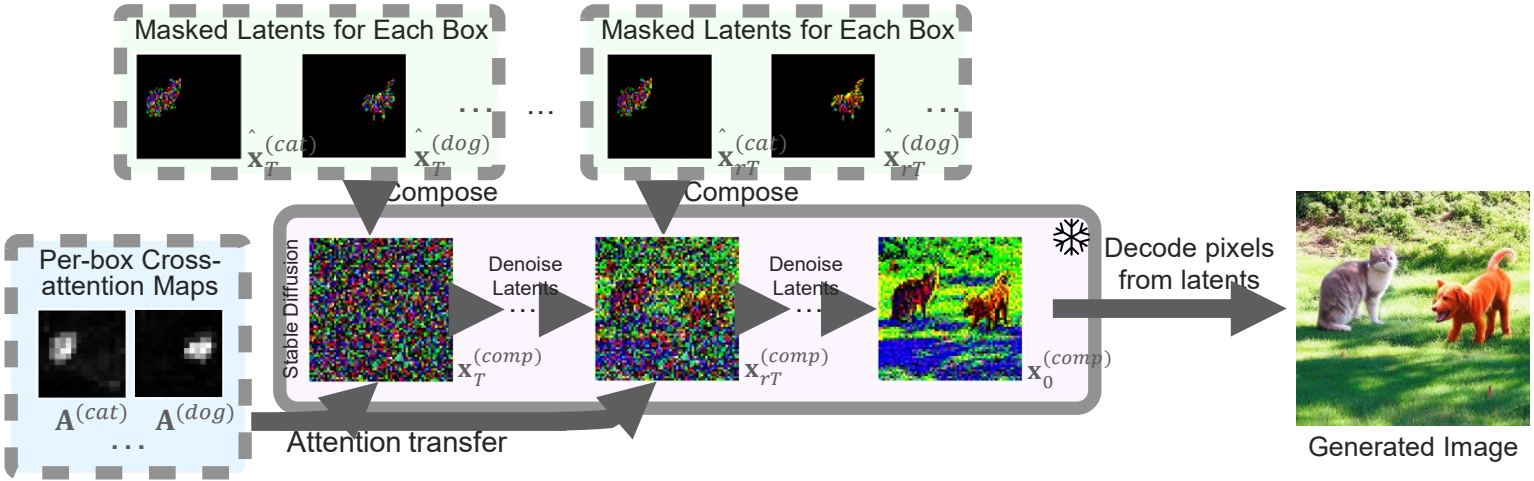
Layout-grounded Stable Diffusion

Stage 2

Generate Masked Latent for Each Box



Compose the Latents for Overall Generation



A gray cat and an orange dog on the grass

Stable Diffusion (Baseline)

Not following the prompt ✗

LMD (Ours)

Accurate attribute binding ✓

A man in red standing next to another woman in blue

Stable Diffusion (Baseline)

Not following the prompt ✗

LMD (Ours)

Accurate attribute binding ✓

A blue cube directly above a red cube with a vase on the left of them

Stable Diffusion (Baseline)

Not following the prompt ✗

LMD (Ours)

Spatial reasoning ✓

An indoor scene with five cats, scattered on the floor

Stable Diffusion (Baseline)

Not following the prompt ✗

LMD (Ours)

Generative numeracy ✓

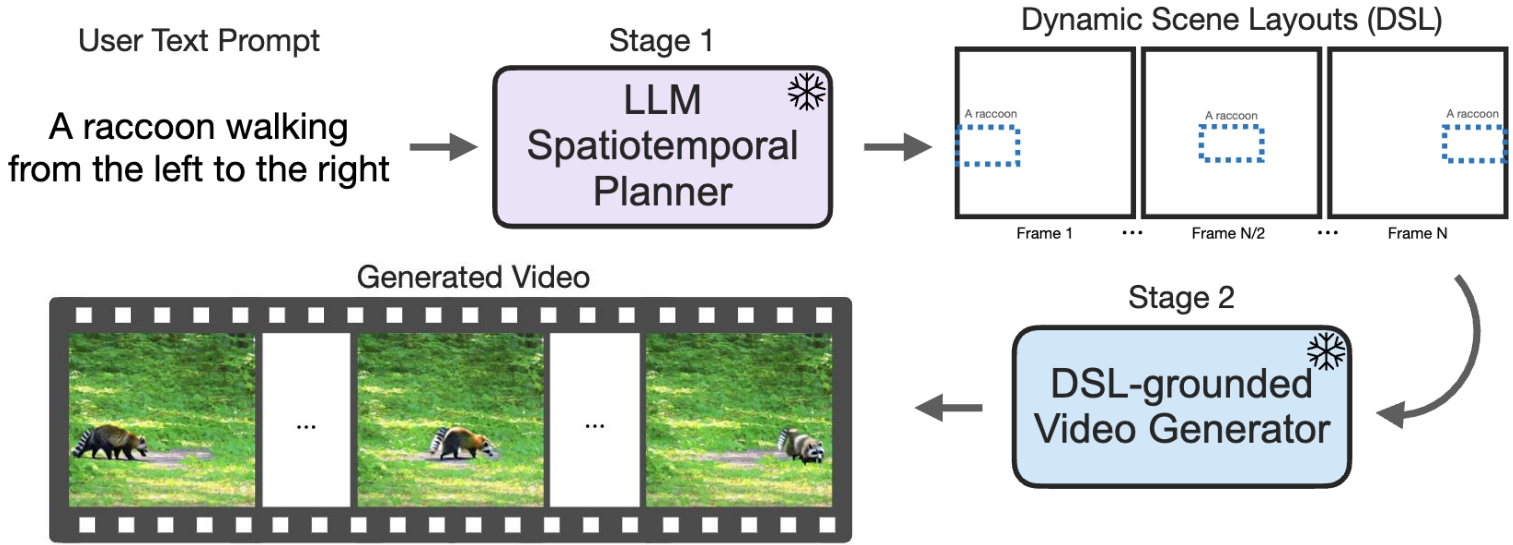
LLMs are able to generate high-quality layouts for text-to-image generation

What about text-to-video generation?

LLM-grounded Video Diffusion Models

Long Lian, Baifeng Shi, Adam Yala, Trevor Darrell, Boyi Li

LLM-grounded Video Diffusion (LVD)



Videos

A raccoon on a wooden barrel floating on a river

ModelScope (Baseline)

Raccoon not on the barrel ✗

LVD (Ours)

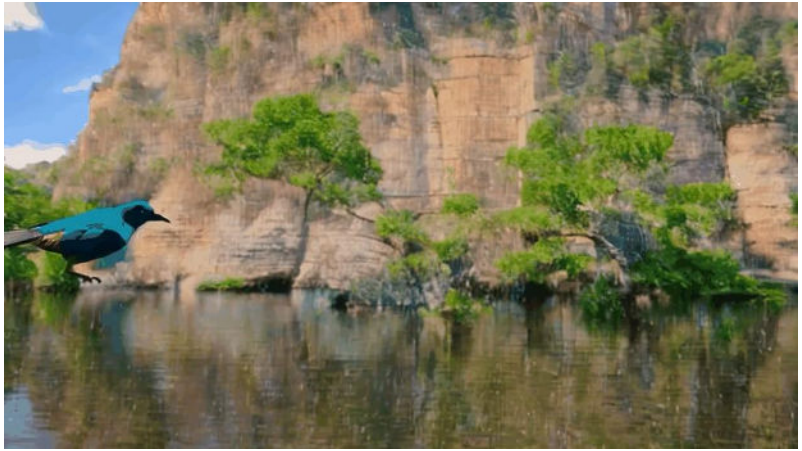
Spatial relationships ✓

Video

A bird flying from the left to the right (of the scene)

ModelScope (Baseline)

Incorrect flying direction ✗



LVD (Ours)

Temporal dynamics ✓

Video

A brown bear dancing with a yellow pikachu

ModelScope (Baseline)

Mixing pikachu and bear ✗

LVD (Ours)

Attribute Binding ✓

What do text only LLMs know?

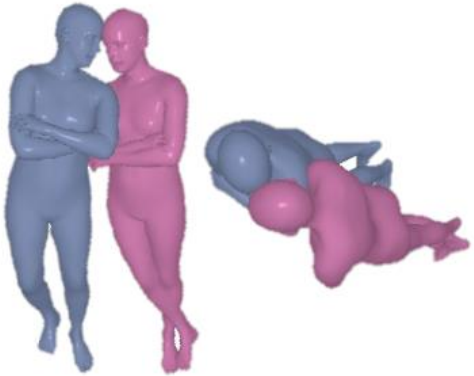
- Perceptual semantics for parsing
- Scene layout for static rendering
- Motion graphs for video generation
- Pose priors for multi-person interaction?

Where to get detailed pose interaction knowledge?

"Their faces are touching as they lean into each other"

from LMM…

… to pose



"The yogi reaches their hands back to touch their heels."

from LMM…

… to pose

Prompt: *"Identify all pairs of body parts of Person 1 and Person 2 that are touching."*

Reference Image: $\boldsymbol{I}$

LMM

Constraint Generation

Arm, Waist (front)
Hand, Waist (front)
Waist(front), Butt
Leg, Butt

```
def lmm_loss(...):
  shoulder_waist_loss = ...
  arm_waist_loss = ...
  hand_stomach_loss = ...
  total_loss = ...
```

$\mathcal{L}_{\text{LMM}}$

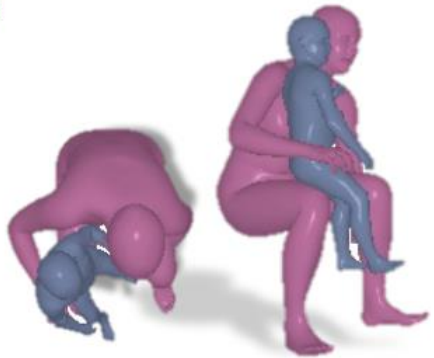
3D Pose Regressor

Initialization: $\boldsymbol{X}$

Constrained Optimization

$$\arg\min(\lambda_{\text{LMM}}\mathcal{L}_{\text{LMM}} + \lambda_{2D}\mathcal{L}_{2D} + \lambda_{\text{GMM}}\mathcal{L}_{\text{GMM}} + \lambda_{\theta}\mathcal{L}_{\theta} + \lambda_{\beta}\mathcal{L}_{\beta})$$

Output: $\boldsymbol{X}'$



Computer Science > Computer Vision and Pattern Recognition

[Submitted on 6 May 2024]

Pose Priors from Language Models

Sanjay Subramanian, Evonne Ng, Lea Müller, Dan Klein, Shiry Ginosar, Trevor Darrell

We present a zero-shot pose optimization method that enforces accurate physical contact constraints when estimating the 3D pose of humans. Our central insight is that since language is often used to describe physical interaction, large pretrained text-based models can act as priors on pose estimation.

We can thus leverage this insight to improve pose estimation by converting natural language descriptors, generated by a large multimodal model (LMM), into tractable losses to constrain the 3D pose optimization. Despite its simplicity, our method produces surprisingly compelling pose reconstructions of people in close contact, correctly capturing the semantics of the social and physical interactions. We demonstrate that our method rivals more complex state-of-the-art approaches that require expensive human annotation of contact points and training specialized models. Moreover, unlike previous approaches, our method provides a unified framework for resolving self-contact and person-to-person contact.

LLMs from Text to Vision and Robotics and back…

- **Are LLMs Grounded? … *Surprisingly so!***
- Reducing VLM Hallucination
- Efficient Scaling of VLMs
- Visual Tokens for Non-linguistic Generation
- Navigation World Models
- 4D Reconstruction for Humanoid Robotics

LLMs from Text to Vision and Robotics and back…

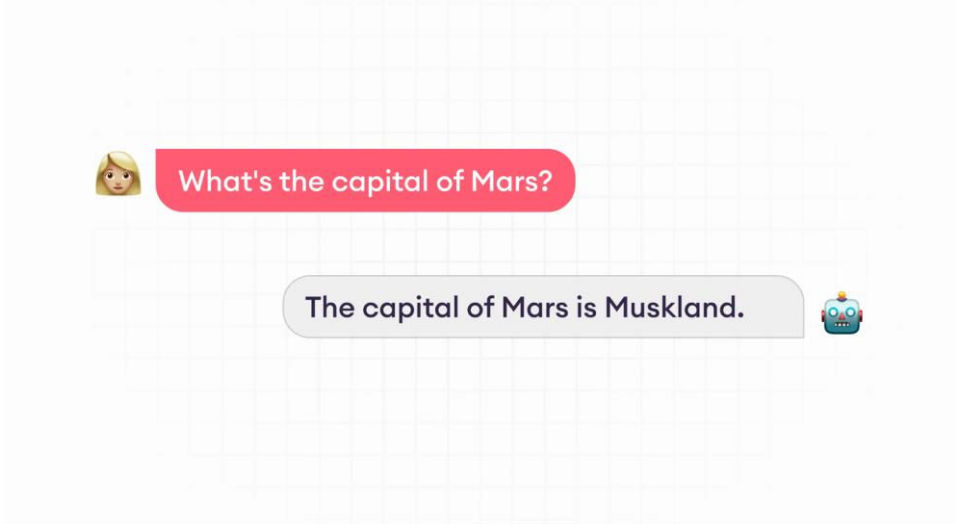
- Are LLMs Grounded? … *Surprisingly so!*
- **Reducing VLM Hallucination**
- Efficient Scaling of VLMs
- Visual Tokens for Non-linguistic Generation
- Navigation World Models
- 4D Reconstruction for Humanoid Robotics

The most prescient problem: LLMs aren’t always grounded.

Untethered -> Unreliable
LLMs without grounding often “fill in the gaps” with convincing but false details.

Why We Should Care
These “hallucinations” break trust and introduce “challenging” errors to catch in downstream tasks

Looking Ahead to Robotic Agents
How can we solve issues in grounding for LLM-based action agents?



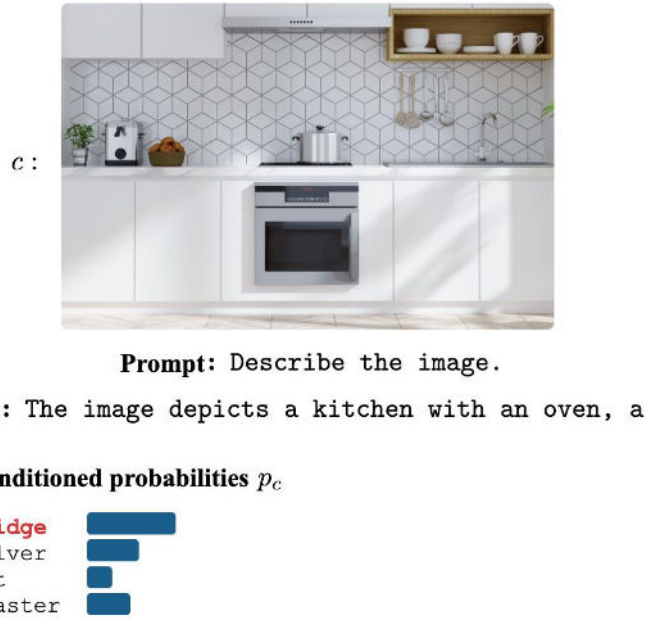
Generate, but Verify: Reducing Visual Hallucination in Vision-Language Models with Retrospective Resampling

Tsung-Han Wu, Heekyung Lee, Jiaxin Ge, Joseph E. Gonzalez, Trevor Darrell, David M. Chan

How can we reduce hallucinations in agentic (and non-agentic) LLMs through retrospective resampling?

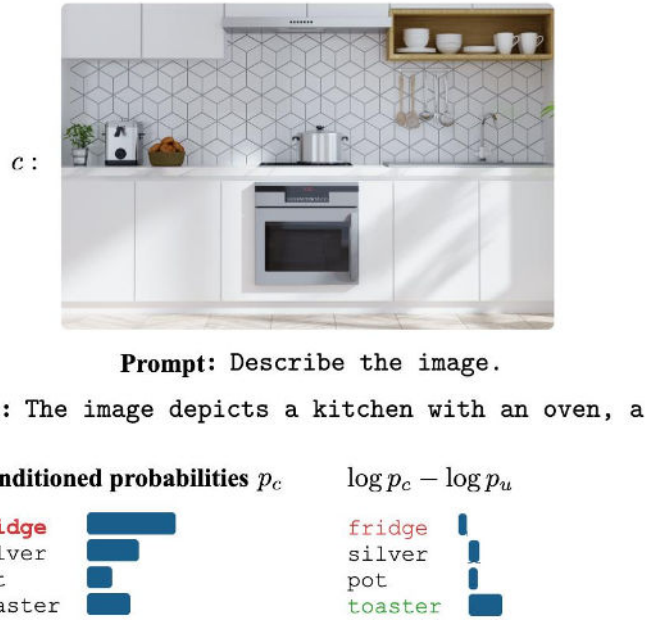
Hallucinations: Not Exactly Failures of Perception

Training data can bias the perception system through the lens of **visual hallucinations**: things that *should* exist in the image, but don't.



Hallucinations: Not Exactly Failures of Perception

Training data can bias the perception system through the lens of **visual hallucinations**: things that *should* exist in the image, but don't.



How do we reduce hallucination now?

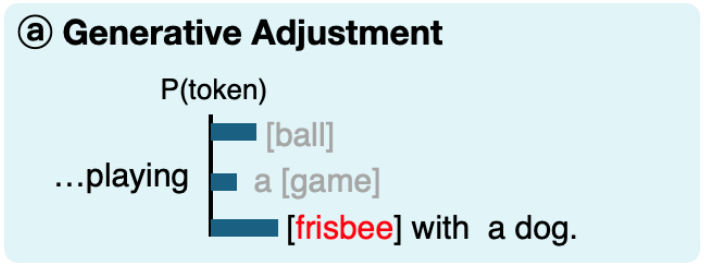
User “Describe this image.”

VLM “The boy is sharing his umbrella…

Existing Paradigm: Generative Adjustment

User “Describe this image.”

VLM “The boy is sharing his umbrella…



Existing Paradigm: Post-Hoc Verification

User "Describe this image."

VLM "The boy is sharing his umbrella…

ⓐ Generative Adjustment

P(token)

…playing [ball] a [game] [frisbee] with a dog.

ⓑ Post-Hoc Verification

…playing frisbee with a dog

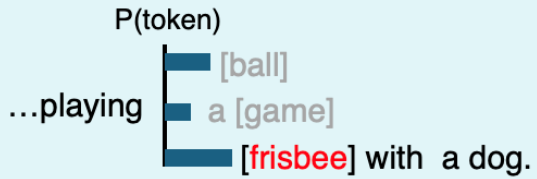
The challenge:

User "Describe this image."

VLM "The boy is sharing his umbrella…

Why not both?

ⓐ Generative Adjustment



ⓑ Post-Hoc Verification

Retrospective thinking

To solve for x in the equation $2x + 5 = 15$, we first subtract 5 from both sides, giving $2x = 10$. Then, dividing by 2, we get $x = 4$. <RETROSPECT> Oh wait, perhaps that was wrong—let me check. The original equation was $2x+5 = 15$, so subtracting 5 gives $2x = 10$, and dividing by 2 CORRECTLY results in $x = 5$. My mistake, the correct answer is $**x = 5$, not 4!**

REVERSE: REtrospective VERification and SElf-correction

What if we could encourage the model to “check itself” for correctness?

User “Describe this image.”

VLM “The boy is sharing his umbrella…

REVERSE: REtrospective VERification and SElf-correction

What if we could encourage the model to "check itself" for correctness?

User "Describe this image."

VLM "The boy is sharing his umbrella…

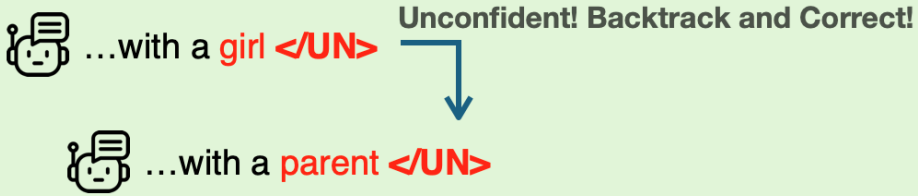
…with a girl **</UN>**

REVERSE: REtrospective VERification and SElf-correction

What if we could encourage the model to "check itself" for correctness?

User "Describe this image."

VLM "The boy is sharing his umbrella…

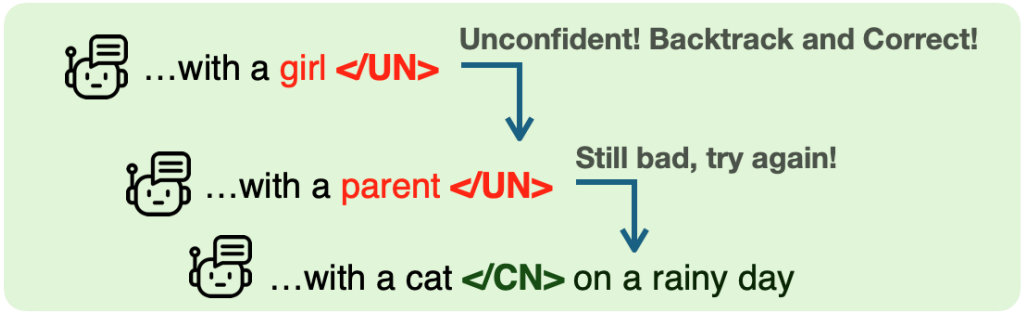


REVERSE: REtrospective VERification and SElf-correction

What if we could encourage the model to "check itself" for correctness?

User "Describe this image."

VLM "The boy is sharing his umbrella…



So how do we get this to work?

What if we could encourage the model to “check itself” for correctness?

Dataset Construction

What if we could encourage the model to “check itself” for correctness?

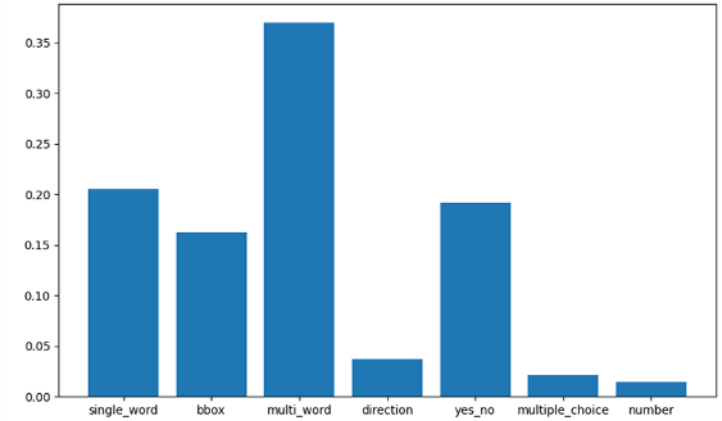
Start with a **600K sample Q/A dataset** for reasoning

Dataset Construction

What if we could encourage the model to “check itself” for correctness?

Start with a **600K sample Q/A dataset** for reasoning

▶ Classify each sample into one of seven categories using rule-based methods.



Dataset Construction

What if we could encourage the model to “check itself” for correctness?

Start with a **600K sample Q/A dataset** for reasoning

- Classify each sample into one of seven categories using rule-based methods.
- Use classical POS tagging to find noun-phrases

Human: “What feature can be seen on the back of the bus?”

GPT: "<SPAN>The back</CN> <SPAN>of the bus</CN> features <SPAN>an advertisement</CN> ."

Dataset Construction

What if we could encourage the model to “check itself” for correctness?

Start with a **600K sample Q/A dataset** for reasoning

- ▶ Classify each sample into one of seven categories using rule-based methods.
- ▶ Use classical POS tagging to find noun-phrases
- ▶ Leverage LLMs to generate plausible hallucination, and replace with <UN> token

```
template = """Given one provided question-answer pair, please select one of the tagged segments in the "answer" and replace it with an
alternative that captures a similar aspect but differs in meaning … Please provide your answer in the following JSON format:
```json
{{
    "Reasoning": "Provide an explanation of why a specific word or phrase was chosen for substitution and the rationale behind the
chosen alternative.",
    "Output": ["Original Text", "Alternative"]
}}
```
Here's the input:
- Question {question}
- Answer: {answer}
"""
```

Dataset Construction

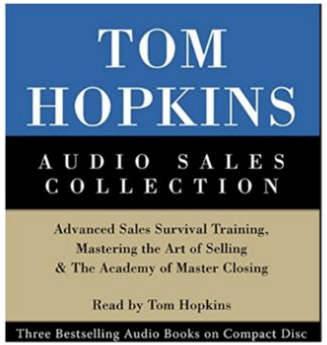
What if we could encourage the model to “check itself” for correctness?

Start with a **600K sample Q/A dataset** for reasoning

- Classify each sample into one of seven categories using rule-based methods.
- Use classical POS tagging to find noun-phrases
- Leverage LLMs to generate plausible hallucination, and replace with <UN> token

Human: “What feature can be seen on the back of the bus?”
GPT: "<SPAN>The back</CN> <SPAN>of the bus</CN> features <SPAN>a window</UN> ."

Dataset Construction

	COCO/train2017/	Captioning Question	VQA task
Image			
Question	"How many total baseball players are shown in the image?"	"Describe this image in your own words."	"Who wrote this book?"
Pos Answer	"There are <SPAN>three baseball players</CN> shown <SPAN>in the image</CN> .",	"The image features <SPAN>an old military aircraft</CN> <SPAN>on display</CN> ..."	"<SPAN>Tom Hopkins</CN>"
Neg Answer	"There are <SPAN>five soccer players</UN>"	"The image features <SPAN>a modern commercial airplane</UN>"	"<SPAN>John Steinbeck</UN>"

Number Attribute Object

Model Fine-Tuning

Model should be trained on <CN> data, but not on <UN> data, since this is **incorrect.**

<SPAN>The back</CN> <SPAN>of the bus</CN> features <SPAN>a window</UN>

Model trained to predict these tokens

Model **ignores** these tokens during training

Model Evaluation

When evaluating the model, monitor the probability of **</UN>** and take action when it exceeds a threshold.

Model Evaluation

When evaluating the model, monitor the probability of **</UN>** and take action when it exceeds a threshold.

<SPAN>The back</CN>

<UN>: 0.0 <UN>: 0.0 <UN>: 0.0 <UN>: 0.0

Model Evaluation

When evaluating the model, monitor the probability of **</UN>** and take action when it exceeds a threshold.

<SPAN>The back</CN> <SPAN>of the bus</CN> features

<UN>: 0.0 <UN>: 0.0 <UN>: 0.0 <UN>: 0.0 <UN>: 0.0 <UN>: 0.0 <UN>: 0.0 <UN>: 0.0 <UN>: 0.0 <UN>: 0.0

Model Evaluation

When evaluating the model, monitor the probability of **</UN>** and take action when it exceeds a threshold.

<SPAN>The back</CN> <SPAN>of the bus</CN> features <SPAN>a window

<UN>: 0.0 <UN>: 0.0 <UN>: 0.0 <UN>: 0.0 <UN>: 0.0 <UN>: 0.0 <UN>: 0.0 <UN>: 0.0 <UN>: 0.0 <UN>: 0.0 <UN>: 0.0 <UN>: 0.0 <UN>: 0.783

Model Evaluation

When evaluating the model, monitor the probability of **</UN>** and take action when it exceeds a threshold.

<SPAN>The back</CN> <SPAN>of the bus</CN> features <SPAN>

<UN>: 0.0 <UN>: 0.0 <UN>: 0.0 <UN>: 0.0 <UN>: 0.0 <UN>: 0.0 <UN>: 0.0 <UN>: 0.0 <UN>: 0.0 <UN>: 0.0 <UN>: 0.0

Model Evaluation

When evaluating the model, monitor the probability of **</UN>** and take action when it exceeds a threshold.

<SPAN>The back</CN> <SPAN>of the bus</CN> features <SPAN> an advertisement.

<UN>: 0.0 <UN>: 0.0 <UN>: 0.0 <UN>: 0.0 <UN>: 0.0 <UN>: 0.0 <UN>: 0.0 <UN>: 0.0 <UN>: 0.0 <UN>: 0.0 <UN>: 0.0 <UN>: 0.0 <UN>: 0.0

Model Evaluation

How do we prevent LLMs from saying the same thing again?

Rejection Sampling

User: <image> Describe this image. Assistant: …</CN>

Temperature Updates: $T = T + \Delta T$

Model Evaluation

How do we prevent LLMs from saying the same thing again?

Rejection Sampling

User: <image> Describe this image. Assistant: …</CN>

Temperature Updates: $T = T + \Delta T$

Or

Query Rewriting

User: <image> Describe this image. (Hint: potential incorrect phrases →a frisbee) Assistant: …</CN>

Preliminary Results: Generative Tasks

Hallucination Words: ['sun', 'bird', 'dog', 'man'];
Safe Words: ['grass', 'woman', 'football', 'court']

Woodpecker --> Hall: frisbee

The image features a woman([0.067, 0.051, 0.453, 0.938]) playing frisbee on a grassy field([0.001, 0.002, 0.999, 0.998]). There is no soccer ball or soccer in the image. There is no ball or person in the image.

REVERSE --> Hall:

The image features a young woman wearing a black and white striped shirt, running through a lush green field. She is in the process of kicking a soccer ball with her leg, which is positioned towards the right side of the field. The woman appears to be enjoying her time playing soccer in the
open space. The field is surrounded by grass, creating a natural and inviting environment for the game.

OPERA --> Hall: bench

The image features a young woman playing soccer on a grassy field. She is running towards a soccer ball, which is positioned towards the right side of the field. The woman appears to be focused and determined as she approaches the ball. In the background, there is a bench located on the left side
of the field, possibly for players to rest or for spectators to sit and watch the game.

LLaVA --> Hall:

The image features a young woman playing soccer on a grassy field. She is running towards a soccer ball, which is positioned towards the right side of the field. The woman appears to be focused and determined as she approaches the ball. In the background, there are two other people, one on the left side and another on the right side of the field. The scene captures the excitement and energy of a soccer game.

Preliminary Results: Generative Tasks

```
Hallucination Words: ['sun', 'sky', 'cloud',
'flower', 'person']; Safe Words: ['forest',
'river', 'grass', 'bicycle', 'toy', 'cat', 'road',
'sky']

**Woodpecker** --> Hall: basket

The image features a red bicycle([0.006, 0.043,
0.998, 0.998]) parked on a grassy area([0.0, 0.613,
0.999, 0.999]) next to a body of water([0.0, 0.451,
0.998, 0.848]). The bicycle is equipped with a
basket, and a small cat figurine([0.33, 0.259,
0.458, 0.5]) is placed on the handlebars([0.058,
0.042, 0.829, 0.556]). The water is visible in the
background, creating a serene and peaceful
atmosphere.

**REVERSE** --> Hall: child

The image features a red bicycle parked near a body
of water, possibly a river or a lake. The
bicycle is positioned on the grass, and a cat is
sitting on the handlebars, adding a playful touch
to the scene. The cat appears to be a stuffed toy,
giving the impression that the bicycle is a toy
for children. The bicycle is a red color, and the
cat is a white color, making the scene visually
appealing and eye- catching.

**OPERA** --> Hall: bench, bench, people

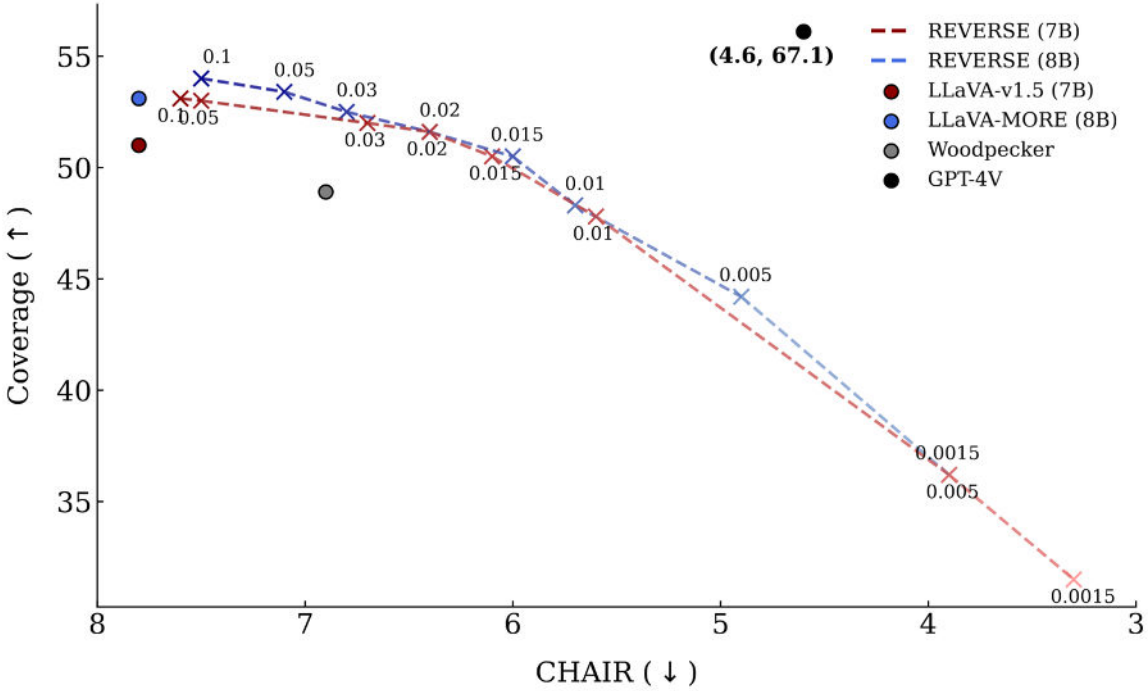
The image features a red bicycle parked next to a
body of water, possibly a river or a lake. On the
handlebars of the b.i.k.e. bike, there is a small
cat figurine or toy, adding a playful touch to the
scene. The bicycle is positioned in such a way that
it appears to be ridden by the cat figurine. In
the background, there are trees and a bench,
creating a serene and peaceful atmosphere. The
bench is
located near the water, providing a place for
people to sit and enjoy the view.

**LLaVA** --> Hall: basket, bench

The image features a red bicycle parked next to a
body of water, possibly a river or a lake. The
bicycle is equipped with a basket, and a small cat
figurine is placed on the handlebars, adding a
playful touch to the scene. In the background,
there are trees and a bench, creating a serene and
peaceful atmosphere. The cat figurine on the
handlebars is the main focal point of the image,
drawing attention to the bicycle and its unique
decoration.
```

Tunable Thresholds for Perception

By tuning the threshold at which we re-sample the outputs, we can **trade-off between expressivity and groundedness**



For more info...

Generate, but Verify: Reducing Visual Hallucination in Vision-Language Models with Retrospective Resampling

Tsung-Han Wu, Heekyung Lee, Jiaxin Ge, Joseph E. Gonzalez, Trevor Darrell, David M. Chan

Project Page

https://reverse-vlm.github.io
Github: https://github.com/tsunghan-wu/reverse_vlm
Huggingface: https://huggingface.co/tsunghanwu/reverse_llava_v15

LLMs from Text to Vision and Robotics and back…

- Are LLMs Grounded? … *Surprisingly so!*
- Reducing VLM Hallucination **with REVERSE Retrospective Sampling**
- **Efficient Scaling of VLMs**
- Visual Tokens for Non-linguistic Generation
- Navigation World Models
- 4D Reconstruction for Humanoid Robotics

Scaling Vision Pre-Training to 4K Resolution


Baifeng Shi


Boyi Li


Han Cai


Yao Lu


Sifei Liu


Marco Pavone


Jan Kautz


Song Han

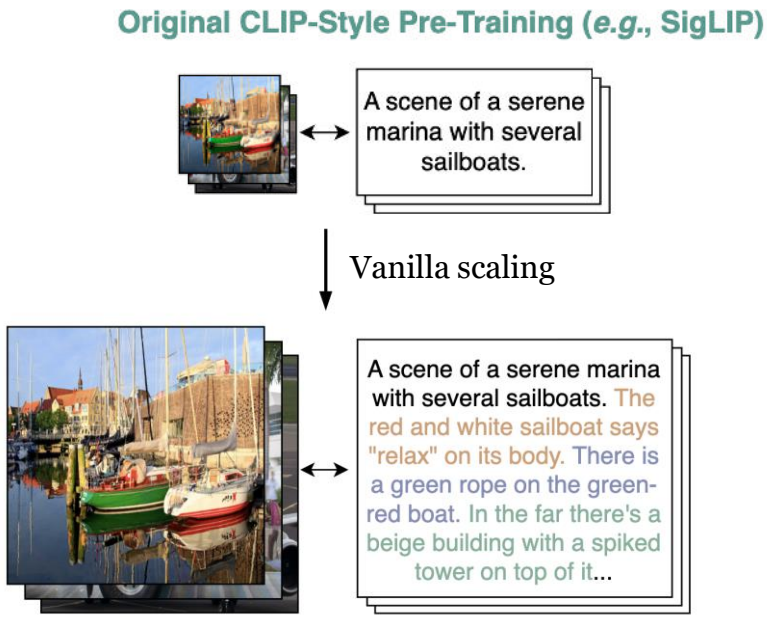

Trevor Darrell


Pavlo Molchanov


Hongxu Yin

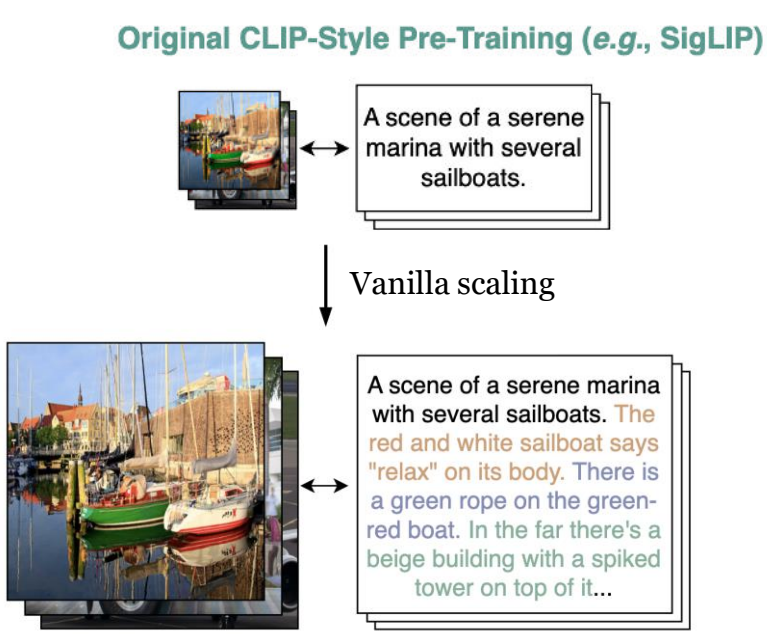
UC Berkeley NVIDIA

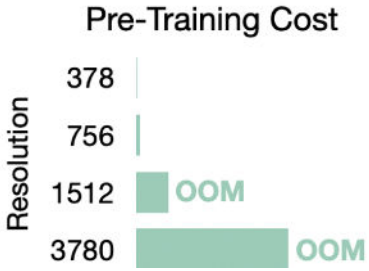
Why couldn’t we do it before?



Scaling Vision Pre-Training to 4K Resolution, Shi et al., CVPR 2025 (highlight)

Why couldn’t we do it before?

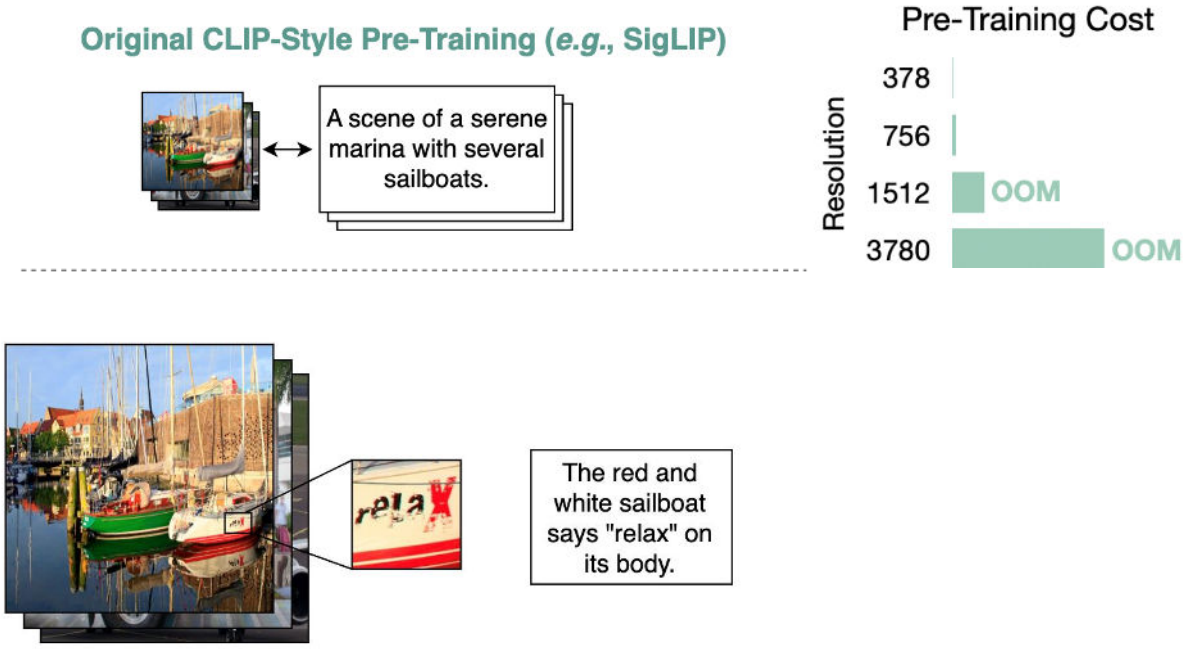




Processing the whole image is slow

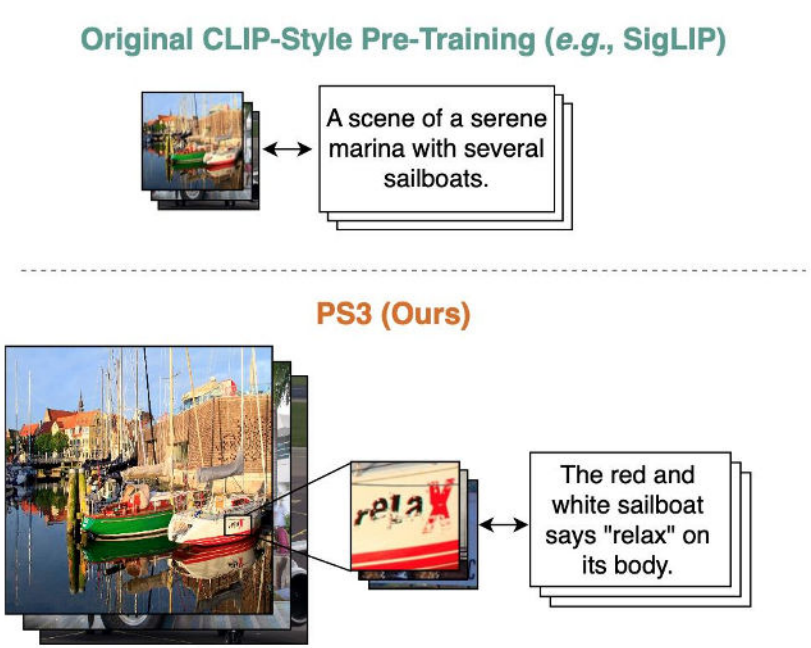
Scaling Vision Pre-Training to 4K Resolution, Shi et al., CVPR 2025 (highlight)

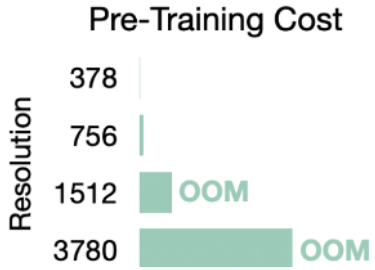
Does it really need to look at everywhere?



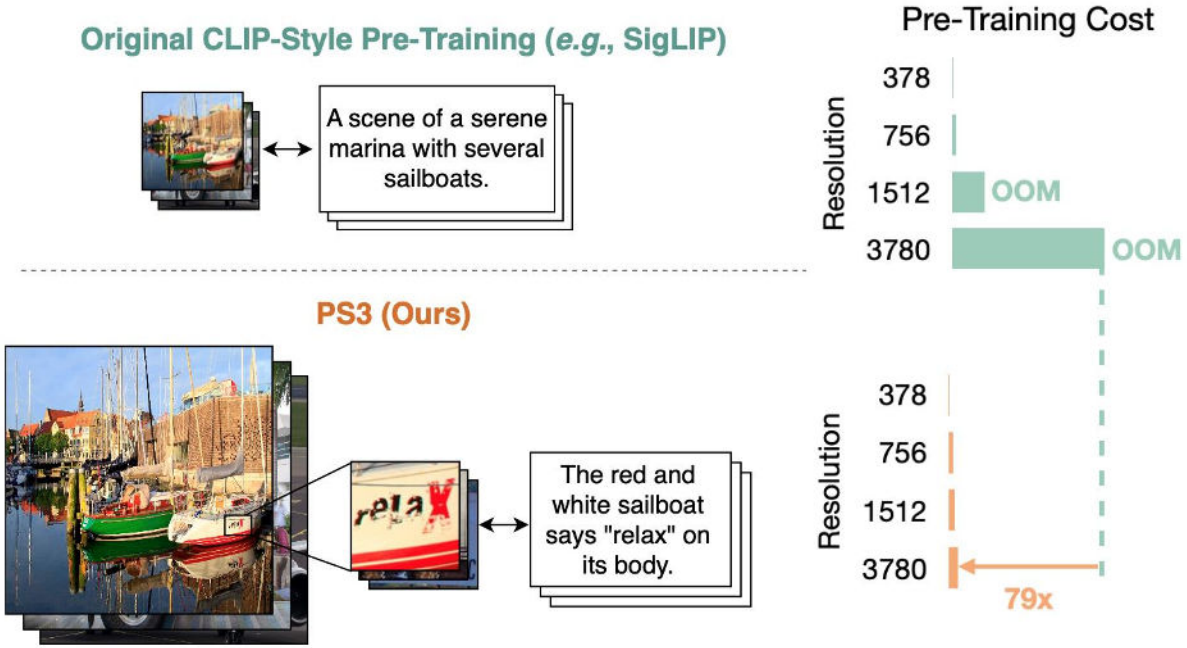
Scaling Vision Pre-Training to 4K Resolution, Shi et al., CVPR 2025 (highlight)

PS3: Localized Contrastive Learning





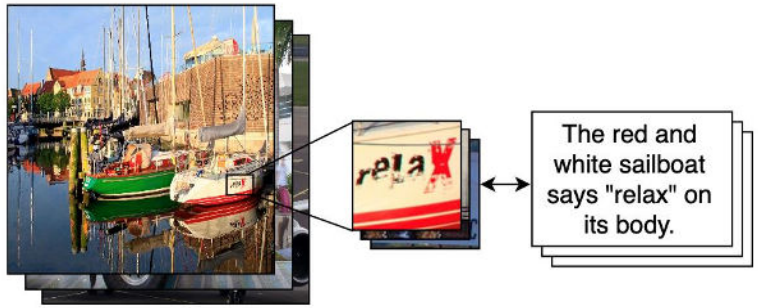
Scaling to 4K Res With Near-Constant Cost



Scaling Vision Pre-Training to 4K Resolution, Shi et al., CVPR 2025 (highlight)

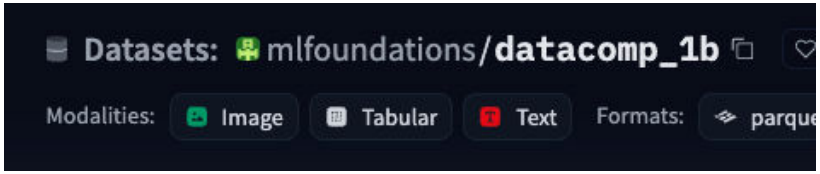
The Recipe

The Recipe



- Data
  - High-res images
  - Pairs of local regions and local detailed captions

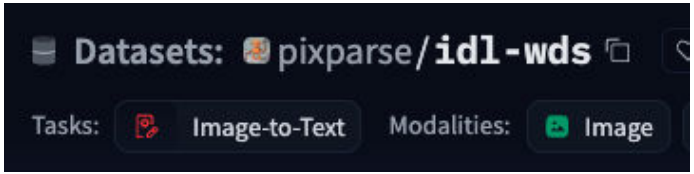
Data – High-Res Images

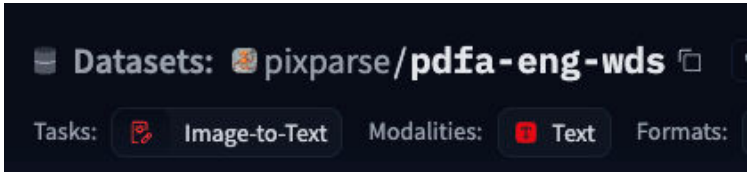


APRIL 5, 2023

SA-1B Dataset

Segment Anything 1 Billion (SA-1B) is a dataset designed for training general-purpose object segmentation models from open world images. The dataset was introduced in our paper “Segment Anything”.

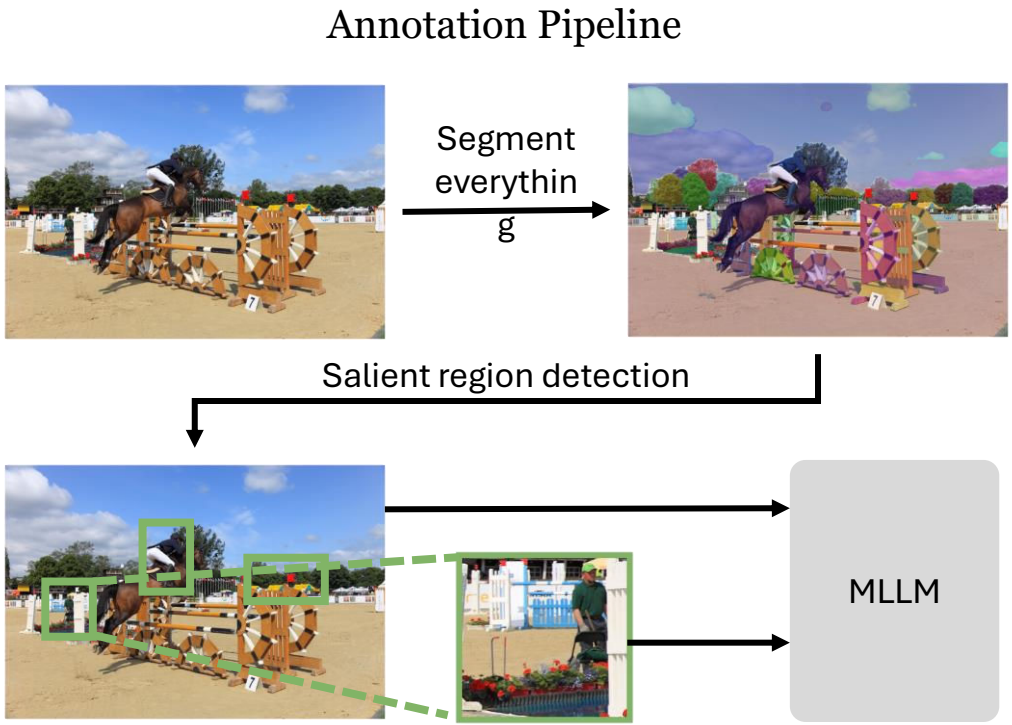




75M images, up to 4K Resolution

Scaling Vision Pre-Training to 4K Resolution, Shi et al., CVPR 2025 (highlight)

Data – Region-Caption Pairs



Scaling Vision Pre-Training to 4K Resolution, Shi et al., CVPR 2025 (highlight)

Data – Region-Caption Pairs

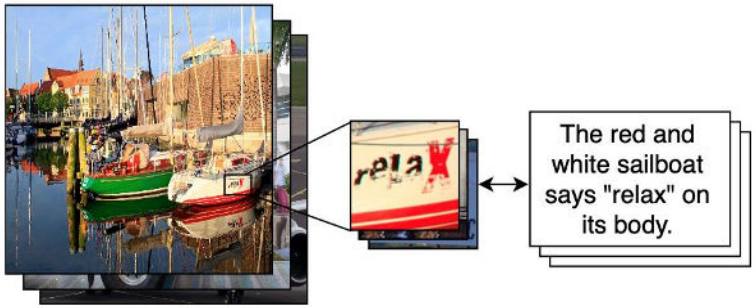
The image shows a vibrant green and orange building with a staircase leading up to a balcony. The building has a door and several lifebuoys hanging on the wall. There are two people sitting on a bench in front of the building, and a bicycle is parked nearby. The text on the building reads "20000".

The image shows a motorcycle parked on a platform with an orange railing. The motorcycle is blue with a black helmet placed on its seat. The background features a green tree and some bushes. The platform appears to be part of a larger structure, possibly a dock or a pier, with a green and orange railing.

The second image shows a portion of the first image, focusing on the area with the boat and the green and orange structure. The boat is a large, colorful vessel with a cabin and a deck, and it is positioned on the water. The green and orange structure appears to be a part of the boat, possibly a cabin or a storage area. The background includes some greenery and a few buildings, indicating a coastal or riverine setting.

The second image is a close-up crop from the first image, focusing on a section of the boat. The boat has a black hull with an orange deck and a green structure on top. There are two Thai flags on the boat, and a small stream of water is visible coming out of the boat. The water is brown and there are green plants in the foreground.

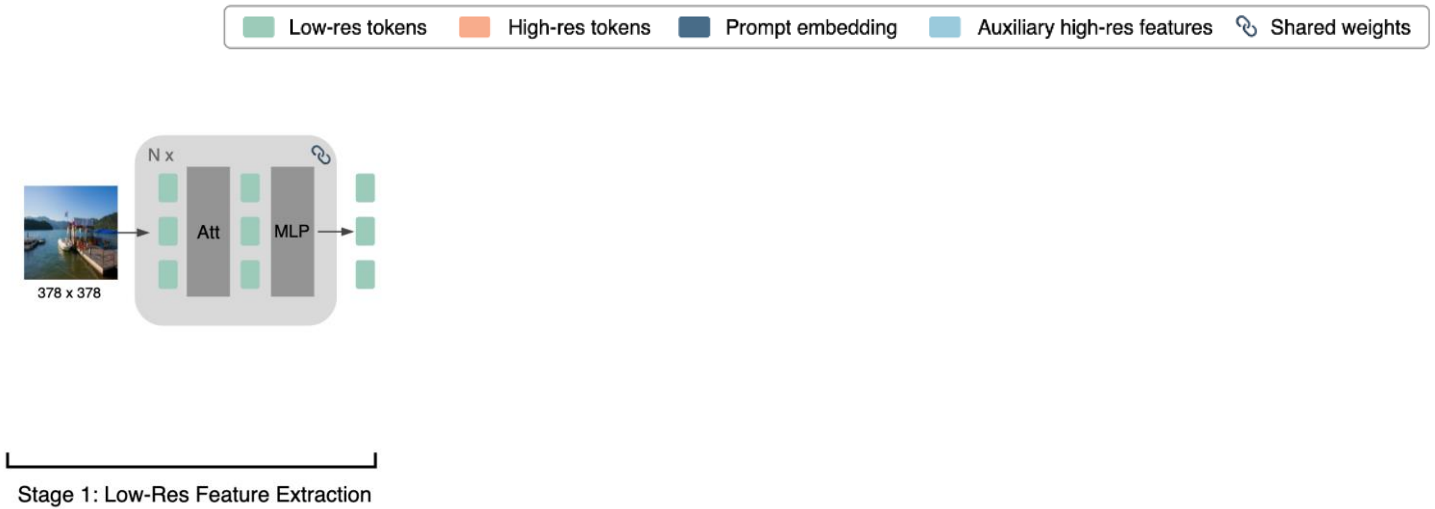
The Recipe



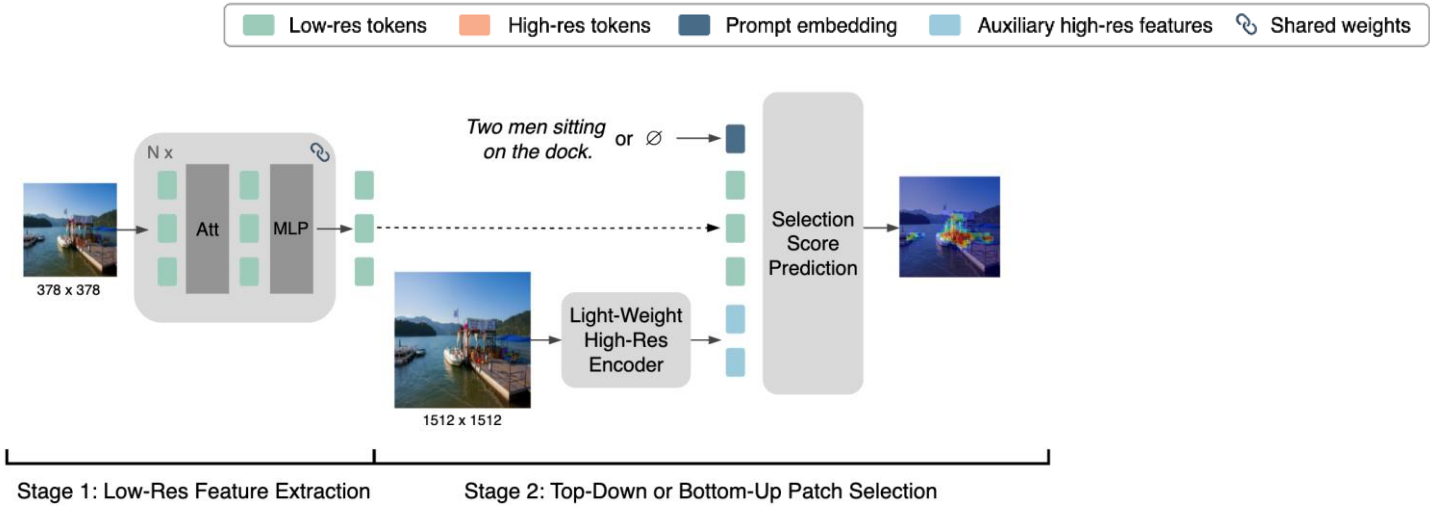
- Data
  - High-res images
  - Pairs of local regions and local detailed captions
- Model
  - Ability of “knowing where to look at”
  - Ability of extracting high-res features from local regions

Scaling Vision Pre-Training to 4K Resolution, Shi et al., CVPR 2025 (highlight)

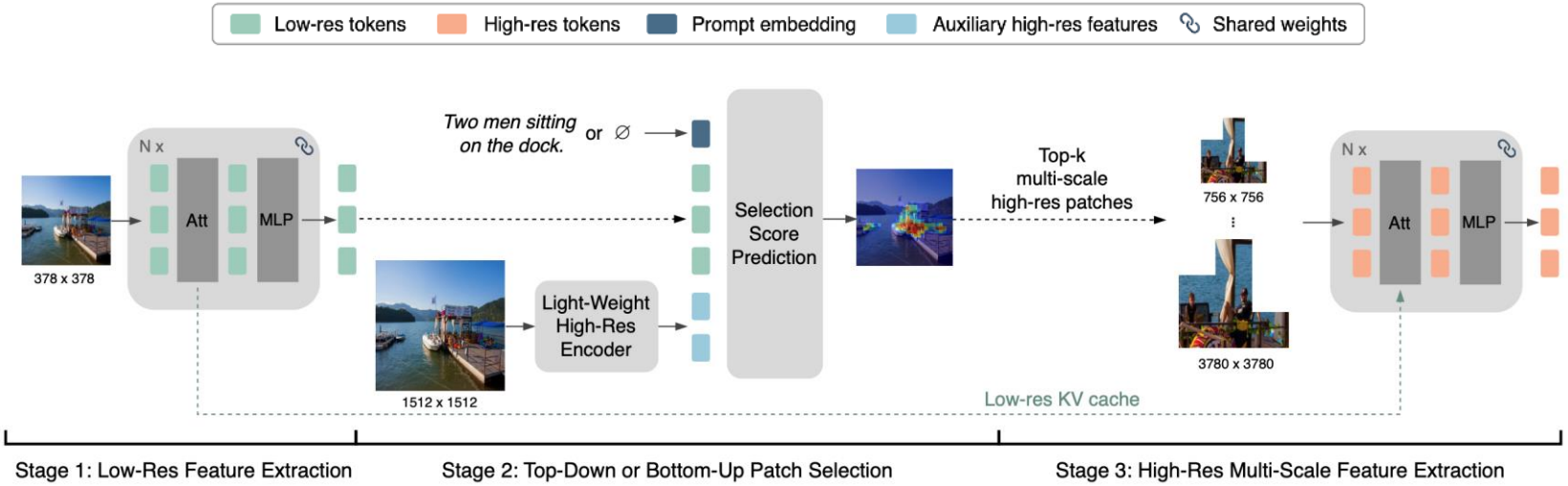
Model


Low-res tokens
High-res tokens
Prompt embedding
Auxiliary high-res features
Shared weights
N x
Att
MLP
378 x 378
Stage 1: Low-Res Feature Extraction

Model

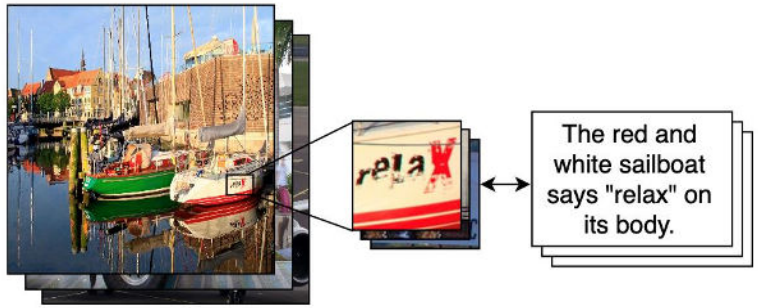

Low-res tokens
High-res tokens
Prompt embedding
Auxiliary high-res features
Shared weights
N x
Att
MLP
378 x 378
Two men sitting on the dock.
or ∅
Light-Weight High-Res Encoder
1512 x 1512
Selection Score Prediction
Stage 1: Low-Res Feature Extraction
Stage 2: Top-Down or Bottom-Up Patch Selection

Model



Scaling Vision Pre-Training to 4K Resolution, Shi et al., CVPR 2025 (highlight)

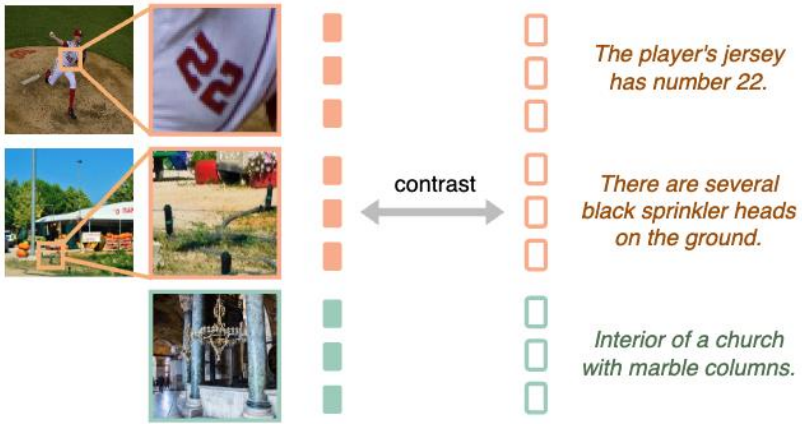
The Recipe



- Data
  - High-res images
  - Pairs of local regions and local detailed captions
- Model
  - Ability of “knowing where to look at”
  - Ability of extracting high-res features from local regions
- Algorithm
  - Optimize region selection
  - Optimize high-res features

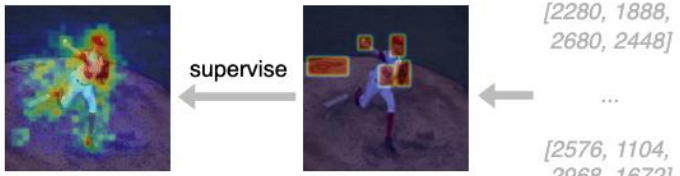
Scaling Vision Pre-Training to 4K Resolution, Shi et al., CVPR 2025 (highlight)

Algorithm



(a) Vision-language contrastive learning

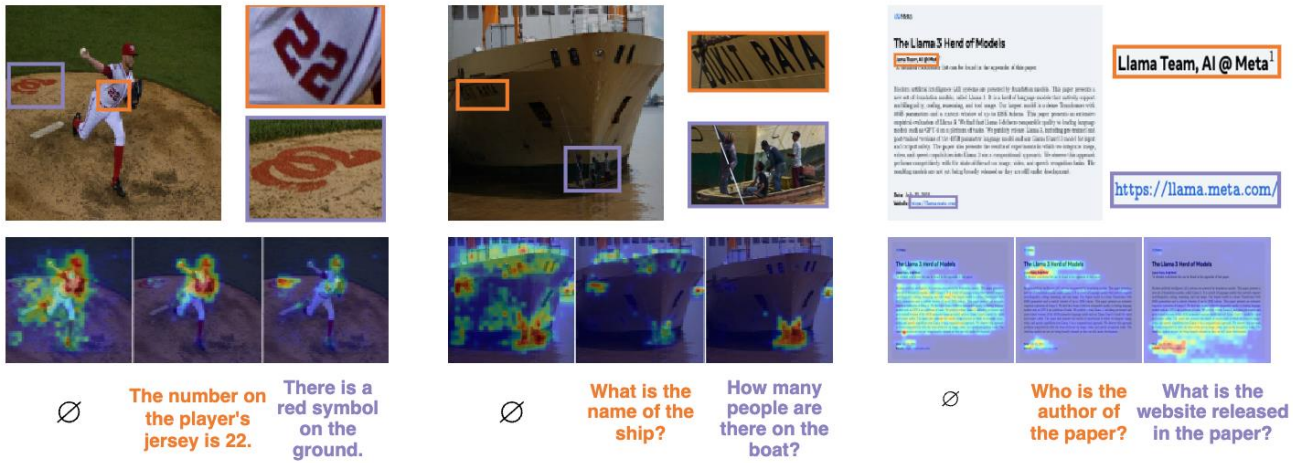
(b) top-down patch selection supervision



(c) bottom-up patch selection supervision

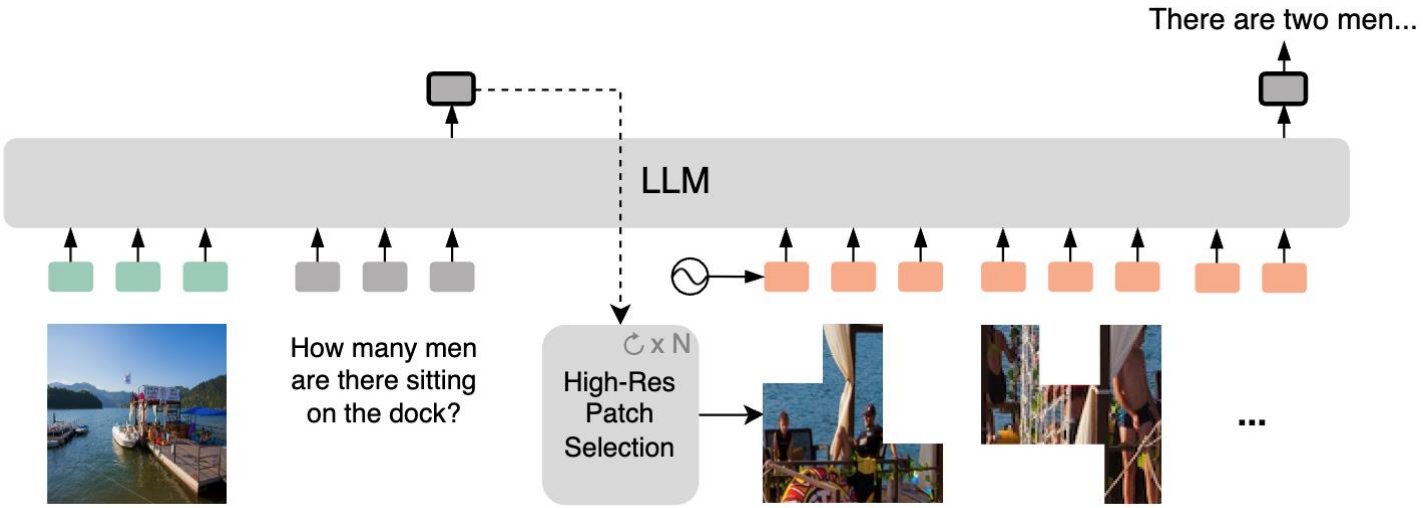
Putting it all together…

PS3 can process up to **4K resolution** and encode **arbitrary high-res regions** based on image saliency or **text prompts**.



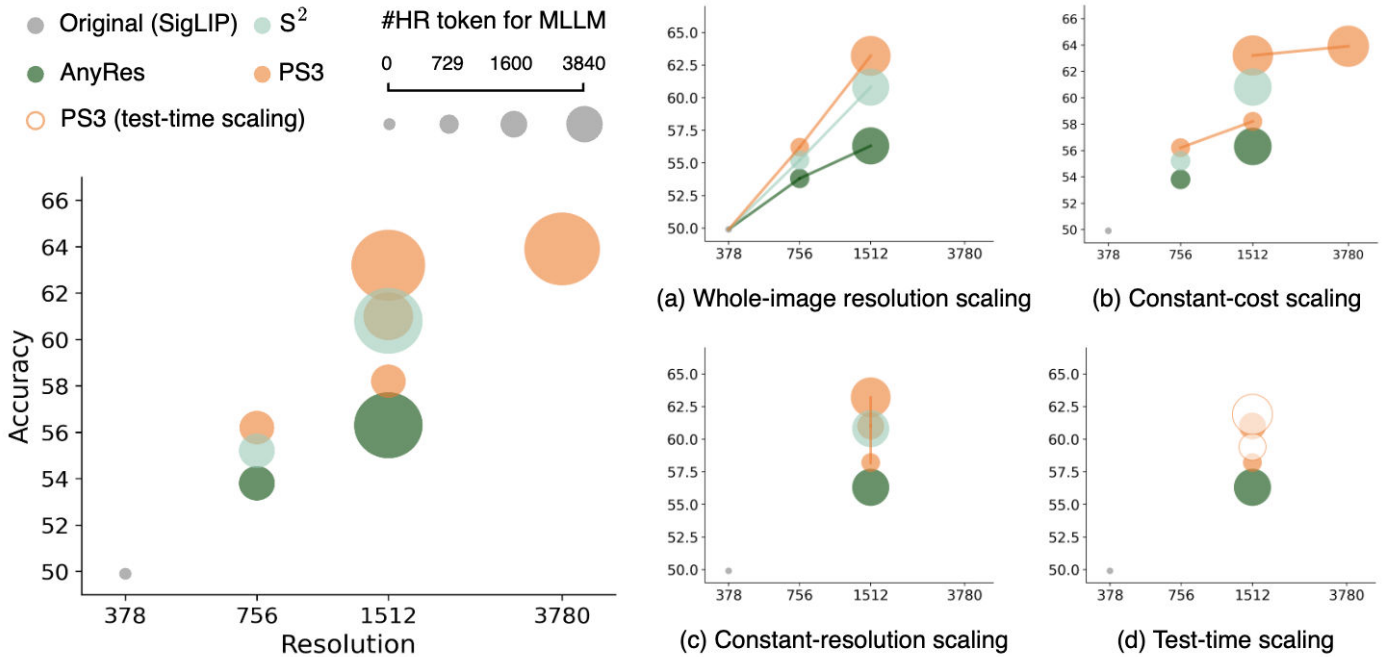
Scaling Vision Pre-Training to 4K Resolution, Shi et al., CVPR 2025 (highlight)

VILA-HD: A High-Res MLLM Built with PS3



Scaling Vision Pre-Training to 4K Resolution, Shi et al., CVPR 2025 (highlight)

Superior Scaling Properties



Scaling Vision Pre-Training to 4K Resolution, Shi et al., CVPR 2025 (highlight)

SOTA Performances

	Res.	Select	#Token	ChartQA (test)	DocVQA (test)	InfoVQA (test)	MathVista (testmini)	MMBench (en-dev)	MMMU-Pro (standard)	OCRBench (test)	V* Bench (test)	RealWorldQA (test)	TextVQA (val)	4KPro (test)
Proprietary														
GPT-4o [40]	-	-	-	85.7	92.8	-	63.8	-	54.0	736	53.7	58.6	-	59.7
Claude 3.5 Sonnet [1]	-	-	-	90.8	95.2	49.7	67.7	-	55.0	788	23.0	59.9	-	29.0
Gemini-1.5-Pro [86]	-	-	-	87.2	93.1	81.0	63.9	-	49.4	754	60.3	70.4	78.7	59.7
Open-source														
VILA-1.5-8B [51]	336	-	576	52.7	40.6	25.9	36.7	68.9	-	-	-	52.7	68.5	33.9
Cambrian-1-8B [88]	1024	-	-	73.3	77.8	-	49.0	75.9	-	624	59.2	64.2	71.7	50.0
NVILA-8B [61]	1552	-	3072	**86.1**	93.7	70.7	65.4	87.6	33.6	794	67.2	66.4	80.1	58.1
MM1.5-7B [111]	2016	-	5184	78.6	88.1	59.5	47.6	-	-	635	-	62.5	76.5	-
LLaVA-OV-7B [48]	2304	-	7252	80.0	87.5	68.8	63.2	80.8	29.5	-	69.2	66.3	-	67.7
IXC2-4KHD [24]	2479	-	7920	81.0	90.0	68.6	57.8	80.2	-	675	-	-	77.2	42.8
IXC-2.5-7B [112]	2743	-	10000	82.2	90.9	70.0	59.6	82.2	-	690	45.6	67.8	78.2	32.3
InternVL2-8B [87]	2833	-	10496	83.3	91.6	74.8	58.3	81.7	32.5	794	65.8	64.4	77.4	58.1
Qwen2-VL-7B [93]	3584	-	16384	83.0	**94.5**	**76.5**	58.2	-	-	**866**	71.0	70.1	**84.3**	71.0
VILA-HD-1.5K-8B	1512	33%	1411	81.3	88.4	58.2	65.3	91.8	35.0	768	67.3	68.4	77.3	50.0
	1512	67%	2626	84.2	91.9	65.3	**66.0**	91.8	**35.1**	776	67.5	68.6	78.0	53.2
	1512	100%	3841	84.3	92.0	67.4	64.6	**92.6**	35.0	782	68.1	68.9	78.4	59.7
VILA-HD-4K-8B	3780	6%	1476	82.2	87.1	57.9	63.9	90.8	34.6	753	68.2	66.5	72.2	62.9
	3780	12%	2756	83.8	91.5	64.5	64.6	91.8	34.7	773	68.8	66.9	77.9	68.8
	3780	18%	4036	84.3	91.7	65.3	64.5	91.8	33.5	774	**71.2**	**70.3**	77.9	**72.6**

Fewer tokens under high resolution

SOTA results on high-res benchmarks

Scaling Vision Pre-Training to 4K Resolution, Shi et al., CVPR 2025 (highlight)

SOTA Efficiency

Heuristic-based token reduction

Method	Select (Test)	ViT Latency	LLM Latency	Text VQA	Chart QA	Doc VQA	Info VQA	OCR Bench	V* Bench	Real World	*Avg*
1512 Resolution											
Full	100%	0.286s	0.375s	78.6	84.1	92.2	68.1	787	67.9	69.8	*77.1*
ToMe [7]	50%	0.286s	0.260s	74.1	70.2	59.7	47.3	622	66.8	67.2	*63.9*
FastV [14]	50%	0.286s	0.264s	**78.2**	81.2	**90.0**	60.4	769	66.2	69.0	*74.6*
VisionZip [106]	50%	0.286s	0.260s	75.2	77.2	79.8	55.7	722	64.0	67.1	*70.2*
PS3	50%	**0.167s**	0.260s	77.7	**83.4**	89.8	**60.8**	**774**	**67.9**	**69.1**	***75.2***
ToMe [7]	25%	0.286s	0.180s	72.5	65.5	51.7	42.8	61.1	62.2	63.4	*59.9*
FastV [14]	25%	0.286s	0.185s	76.1	66.3	78.1	49.5	651	64.6	65.2	*66.6*
VisionZip [106]	25%	0.286s	0.180s	74.6	76.0	72.8	51.5	694	62.7	64.6	*67.4*
PS3	25%	**0.096s**	0.180s	**76.8**	**80.4**	**84.4**	**54.6**	**738**	**65.7**	**67.8**	***71.9***
3780 Resolution											
Full	100%	1.812s	OOM	-	-	-	-	-	-	-	-
ToMe [7]	20%	1.812s	OOM	-	-	-	-	-	-	-	-
FastV [14]	20%	1.812s	OOM	-	-	-	-	-	-	-	-
VisionZip [106]	20%	1.812s	OOM	-	-	-	-	-	-	-	-
PS3	20%	**0.417s**	0.383s	77.8	83.9	91.6	65.0	773	72.8	70.1	*76.9*

Better efficiency and performance

The only one that can process 4K resolution

Scaling Vision Pre-Training to 4K Resolution, Shi et al., CVPR 2025 (highlight)

Question:
Where does the exit lead to?

Qwen2-VL
(resolution: 1792 x 1792)
❌ Latency 3.61s
✅ Accuracy 92%

Qwen2-VL
(resolution: 1024 x 1024)
✅ Latency 1.46s
❌ Accuracy 50%

VILA-HD
(resolution: 4K x 4K)
✅ Latency 1.22s
✅ Accuracy 100%

LLMs from Text to Vision and Robotics and back…

- Are LLMs Grounded? … *Surprisingly so!*
- Reducing VLM Hallucination with REVERSE Retrospective Sampling
- **Efficient Scaling of VLMs to 4K Resolution with via PS3**
- Visual Tokens for Non-linguistic Generation
- Navigation World Models
- 4D Reconstruction for Humanoid Robotics

LLMs from Text to Vision and Robotics and back…

- Are LLMs Grounded? … *Surprisingly so!*
- Reducing VLM Hallucination with REVERSE Retrospective Sampling
- Efficient Scaling of VLMs to 4K Resolution with via PS3
- **Visual Tokens for Non-linguistic Generation**
- Navigation World Models
- 4D Reconstruction for Humanoid Robotics

How to Describe an Image?

Describe the Image with Text

This image features two close-up portraits of cats sitting side by side. Both cats have a short-haired coat with a soft texture, and their expressions are calm and attentive. The cat on the left has a gray coat with faint tabby patterns and piercing yellow eyes. The cat on the right has a slightly lighter, beige-toned coat, also with yellow eyes that are slightly wider, giving it a curious look. Their ears are upright, and their whiskers are prominent. The background is plain white, emphasizing the cats as the main focus of the image. The lighting is soft and evenly distributed, highlighting the details in their fur and facial features.

GPT-4o

110 words

Wang, X., Zhou, X., Fathi, A., Darrell, T. and Schmid, C. “Visual Lexicon: Rich Image Features in Language Space” CVPR 2025.

How to Describe an Image?

Describe the Image with Text

This image features two close-up portraits of cats sitting side by side. Both cats have a short-haired coat with a soft texture, and their expressions are calm and attentive. The cat on the left has a gray coat with faint tabby patterns and piercing yellow eyes. The cat on the right has a slightly lighter, beige-toned coat, also with yellow eyes that are slightly wider, giving it a curious look. Their ears are upright, and their whiskers are prominent. The background is plain white, emphasizing the cats as the main focus of the image. The lighting is soft and evenly distributed, highlighting the details in their fur and facial features.

110 words

Wang, X., Zhou, X., Fathi, A., Darrell, T. and Schmid, C. “Visual Lexicon: Rich Image Features in Language Space” CVPR 2025.

How to Describe an Image?

Describe the Image with Text

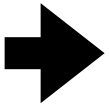
This image features two close-up portraits of cats sitting side by side. Both cats have a short-haired coat with a soft texture, and their expressions are calm and attentive. The cat on the left has a gray coat with faint tabby patterns and piercing yellow eyes. The cat on the right has a slightly lighter, beige-toned coat, also with yellow eyes that are slightly wider, giving it a curious look. Their ears are upright, and their whiskers are prominent. The background is plain white, emphasizing the cats as the main focus of the image. The lighting is soft and evenly distributed, highlighting the details in their fur and facial features.

110 words

Wang, X., Zhou, X., Fathi, A., Darrell, T. and Schmid, C. “Visual Lexicon: Rich Image Features in Language Space” CVPR 2025.

How to Describe an Image?

Describe the Image with Text



Wang, X., Zhou, X., Fathi, A., Darrell, T. and Schmid, C. “Visual Lexicon: Rich Image Features in Language Space” CVPR 2025.

How to Describe an Image?

Describe with Detailed Text

This image features two short-haired cats sitting closely side by side, facing forward. Both cats have alert postures, upright ears, and expressive yellow eyes that capture their calm and curious demeanor.

Detailed Description:

1. **Left Cat:**
   - **Color: The fur is a smooth gray with faint tabby markings visible across the face and body, adding subtle texture to its coat.**
   - **Eyes: Bright yellow with a focused, calm expression. The pupils are moderately dilated, giving it a tranquil but attentive look.**
   - **Ears: Upright and slightly tilted outward, indicating attentiveness.**
   - **Whiskers: Prominent and white, extending outward from the muzzle.**
   - **Pose: The cat's head is slightly tilted, giving it a thoughtful appearance.**
2. **Right Cat:**
   - **Color: The fur has a warm beige tone with softer gradations, appearing slightly lighter than the left cat.**
   - **Eyes: Similarly bright yellow, but with a slightly wider, more curious gaze compared to the left cat.**
   - **Ears: Upright, with a slight inward tilt, creating a sense of focus and alertness.**
   - **Whiskers: Long, white, and delicately curved outward from its face.**
   - **Pose: The head is held level, with an inquisitive expression that contrasts slightly with the calm demeanor of the left cat.**
3. **Background:**
   - The background is plain white, providing a neutral and clean setting that emphasizes the details of the cats without any distractions.
   - Soft, even lighting illuminates the cats, creating gentle highlights on their fur and subtle shadows that add depth and dimension.
4. **Composition:**
   - The cats are positioned close to each other, creating a symmetrical and balanced composition. Their proximity suggests a sense of companionship or familiarity.
   - The focus is sharp, capturing the fine details of the fur, facial features, and whiskers, while the plain background ensures the cats are the central focus.

The image overall conveys a sense of warmth and calm, with both cats displaying slightly different but complementary personal

GPT-4o

302 words

How to Describe an Image?

Describe with Detailed Text

This image features two short-haired cats sitting closely side by side, facing forward. Both cats have alert postures, upright ears, and expressive yellow eyes that capture their calm and curious demeanor.

Detailed Description:

1. **Left Cat:**
   - **Color: The fur is a smooth gray with faint tabby markings visible across the face and body, adding subtle texture to its coat.**
   - **Eyes: Bright yellow with a focused, calm expression. The pupils are moderately dilated, giving it a tranquil but attentive look.**
   - **Ears: Upright and slightly tilted outward, indicating attentiveness.**
   - **Whiskers: Prominent and white, extending outward from the muzzle.**
   - **Pose: The cat's head is slightly tilted, giving it a thoughtful appearance.**
2. **Right Cat:**
   - **Color: The fur has a warm beige tone with softer gradations, appearing slightly lighter than the left cat.**
   - **Eyes: Similarly bright yellow, but with a slightly wider, more curious gaze compared to the left cat.**
   - **Ears: Upright, with a slight inward tilt, creating a sense of focus and alertness.**
   - **Whiskers: Long, white, and delicately curved outward from its face.**
   - **Pose: The head is held level, with an inquisitive expression that contrasts slightly with the calm demeanor of the left cat.**
3. **Background:**
   - The background is plain white, providing a neutral and clean setting that emphasizes the details of the cats without any distractions.
   - Soft, even lighting illuminates the cats, creating gentle highlights on their fur and subtle shadows that add depth and dimension.
4. **Composition:**
   - The cats are positioned close to each other, creating a symmetrical and balanced composition. Their proximity suggests a sense of companionship or familiarity.
   - The focus is sharp, capturing the fine details of the fur, facial features, and whiskers, while the plain background ensures the cats are the central focus.

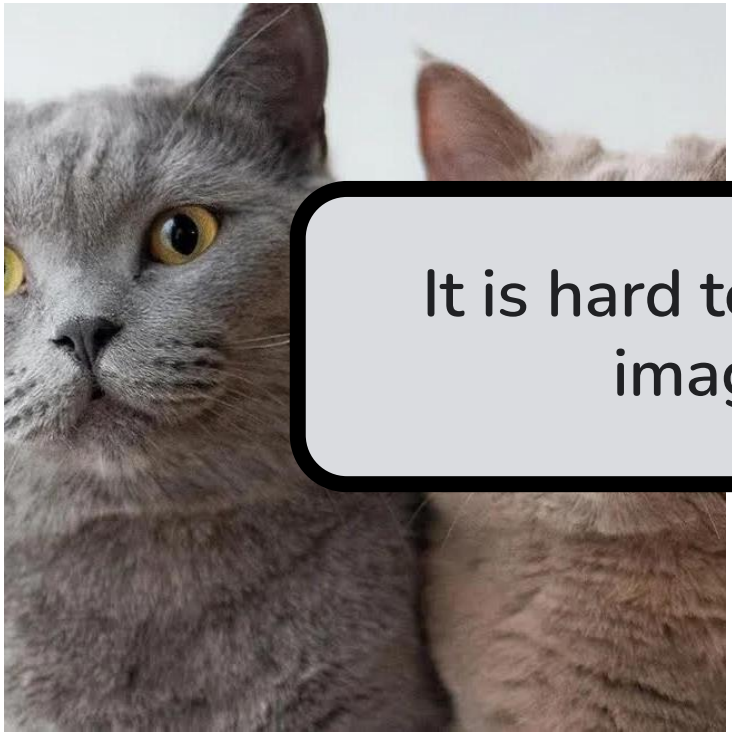
The image overall conveys a sense of warmth and calm, with both cats displaying slightly different but complementary personalities.

302 words

Wang, X., Zhou, X., Fabian, G., Darrell, T. and Schmid, C. “Visual Lexicon: Rich Image Features in Language Space” CVPR 2025.

How to Describe an Image?

Describe with Detailed Text



It is hard to fully describe an image with text

Wang, X., Zhou, X., Fathi, A., Darrell, T. and Schmid, C. “Visual Lexicon: Rich Image Features in Language Space” CVPR 2025.

Text is a user-friendly interface
to interact with models.

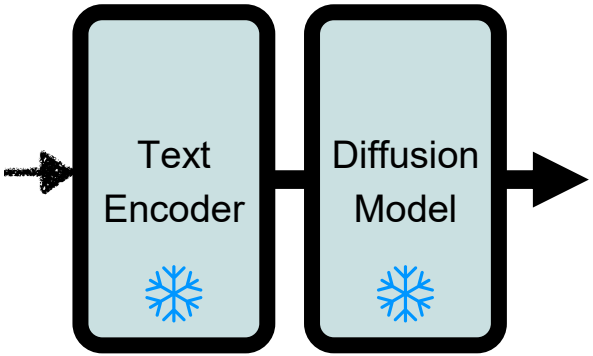
Can we empower linguistic space to capture rich visual details?

ViLex Prompt as a Visual Information Rich Text Prompt

Describe the Image with Text

This image features two close-up portraits of cats sitting side by side. Both cats have a short-haired coat with a soft texture, and their expressions are calm and attentive. The cat on the left has a gray coat with faint tabby patterns and piercing yellow eyes. The cat on the right has a slightly lighter, beige-toned coat, also with yellow eyes that are slightly wider, giving it a curious look. Their ears are upright, and their whiskers are prominent. The background is plain white, emphasizing the cats as the main focus of the image. The lighting is soft and evenly distributed, highlighting the details in their fur and facial features.

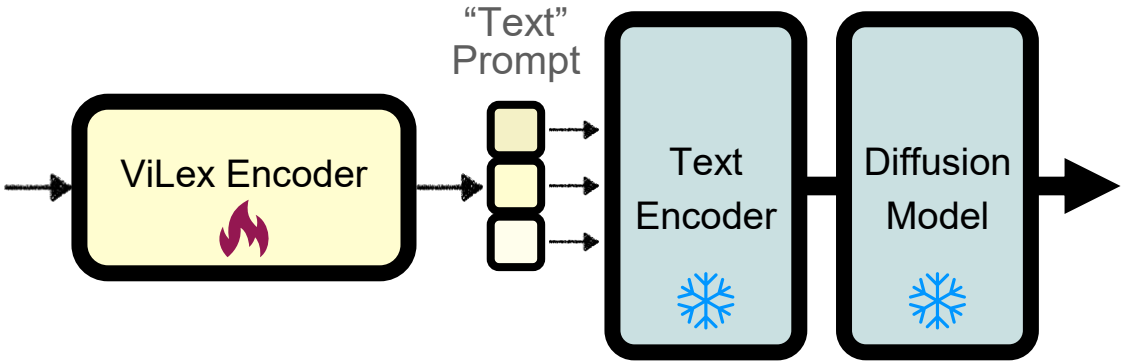
110 words



Wang, X., Zhou, X., Fathi, A., Darrell, T. and Schmid, C. “Visual Lexicon: Rich Image Features in Language Space” CVPR 2025.

ViLex Prompt as a Visual Information Rich Text Prompt

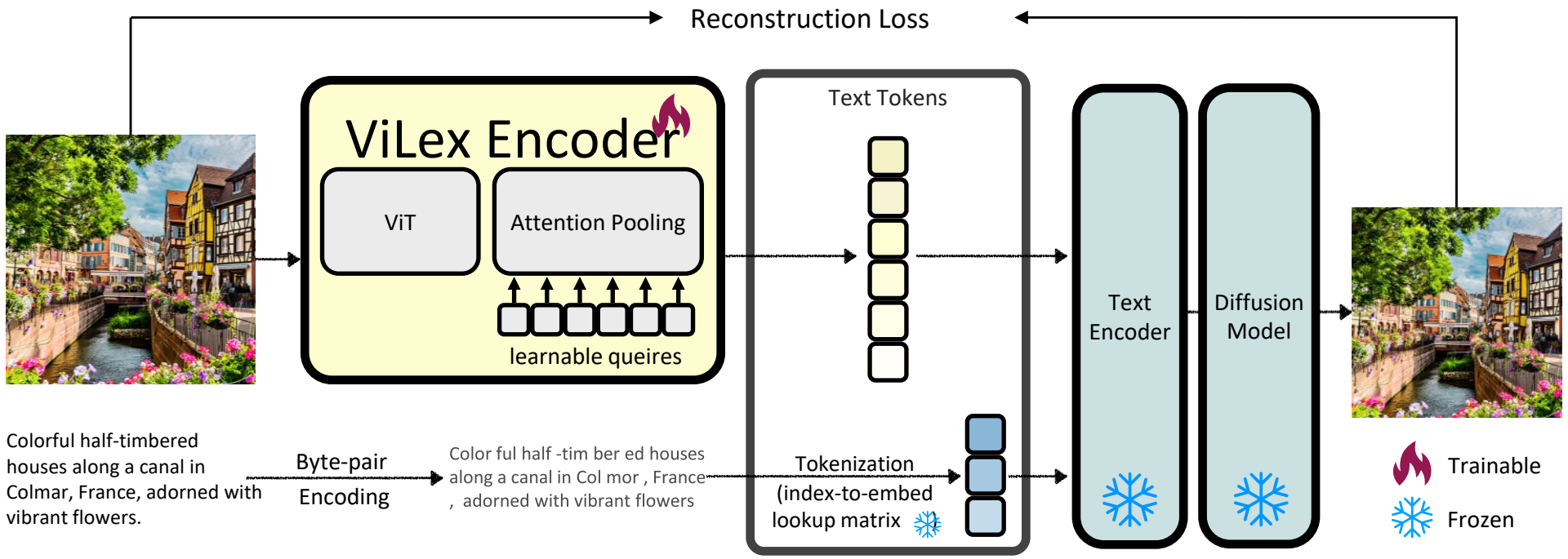
Describe with a Visual Lexicon



Mapping images into the text vocabulary space, effectively creating a new visual "language" that retains intricate visual details

Wang, X., Zhou, X., Fathi, A., Darrell, T. and Schmid, C. "Visual Lexicon: Rich Image Features in Language Space" CVPR 2025.

Learn a Visual Lexicon from Frozen T2I Diffusion Models



ViLex can be trained independently or in conjunction with text prompts.

Wang, X., Zhou, X., Fathi, A., Darrell, T. and Schmid, C. “Visual Lexicon: Rich Image Features in Language Space” CVPR 2025.

ViLex Prompt as a Visual Information Rich Text Prompt

Describe with a Visual Lexicon

Wang, X., Zhou, X., Fathi, A., Darrell, T. and Schmid, C. “Visual Lexicon: Rich Image Features in Language Space” CVPR 2025.

ViLex empower linguistic space to capture visual richness

Linguistic space empowers ViLex to enjoy compositionality

Prompting Text-to-Image Model with [ViLex, Text]

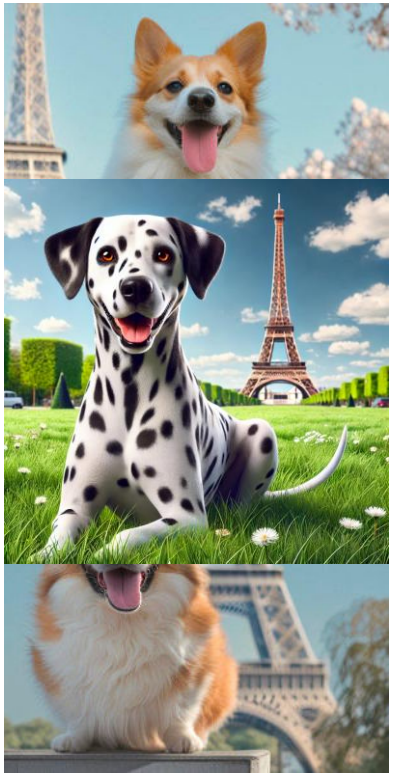
- just like prompting LLMs or VLMs, via embedding images directly into a sentence

ViLex Prompt

Text Prompt

with Eiffel Tower in the background”

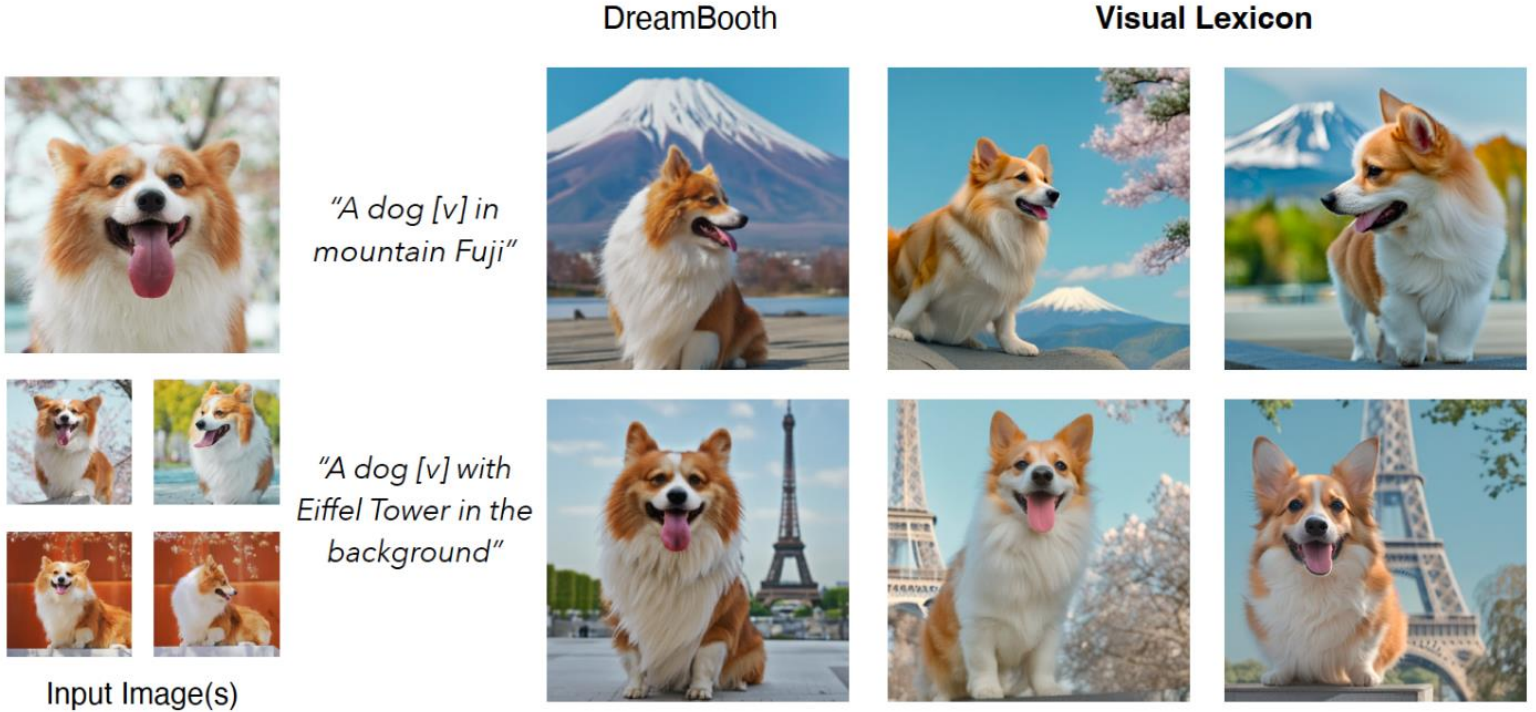
T2I Model



Wang, X., Zhou, X., Fathi, A., Darrell, T. and Schmid, C. “Visual Lexicon: Rich Image Features in Language Space” CVPR 2025.

Zero-shot Unsupervised Image Re-contextualization

Unlike the supervised DreamBooth, ViLex does not need T2I fine-tuning, or modifications to diffusion model architecture.

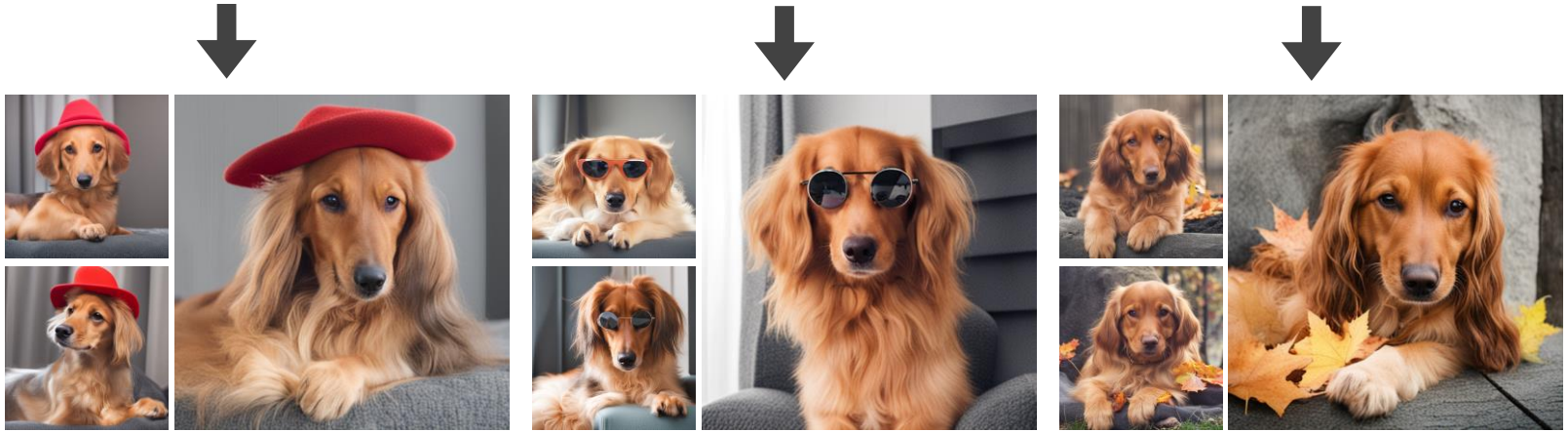


Wang et al. "Visual Lexicon: Rich Image Features in Language Space" CVPR 2025.
Ruiz, et al. DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation. CVPR 2023

Zero-shot Unsupervised Accessorization

ViLex Prompt

Text Prompts: *+ wearing a red hat* *+ wearing sunglasses* *+ autumn leaves*



Wang, X., Zhou, X., Fathi, A., Darrell, T. and Schmid, C. “Visual Lexicon: Rich Image Features in Language Space” CVPR 2025.

Zero-shot Unsupervised Art-Rendition

ViLex Prompt 1

ViLex Prompt 2

ViLex Prompt 3

Text Prompt: ***"An image of [ViLex Prompt] in Vincent Van Gogh's Style"***

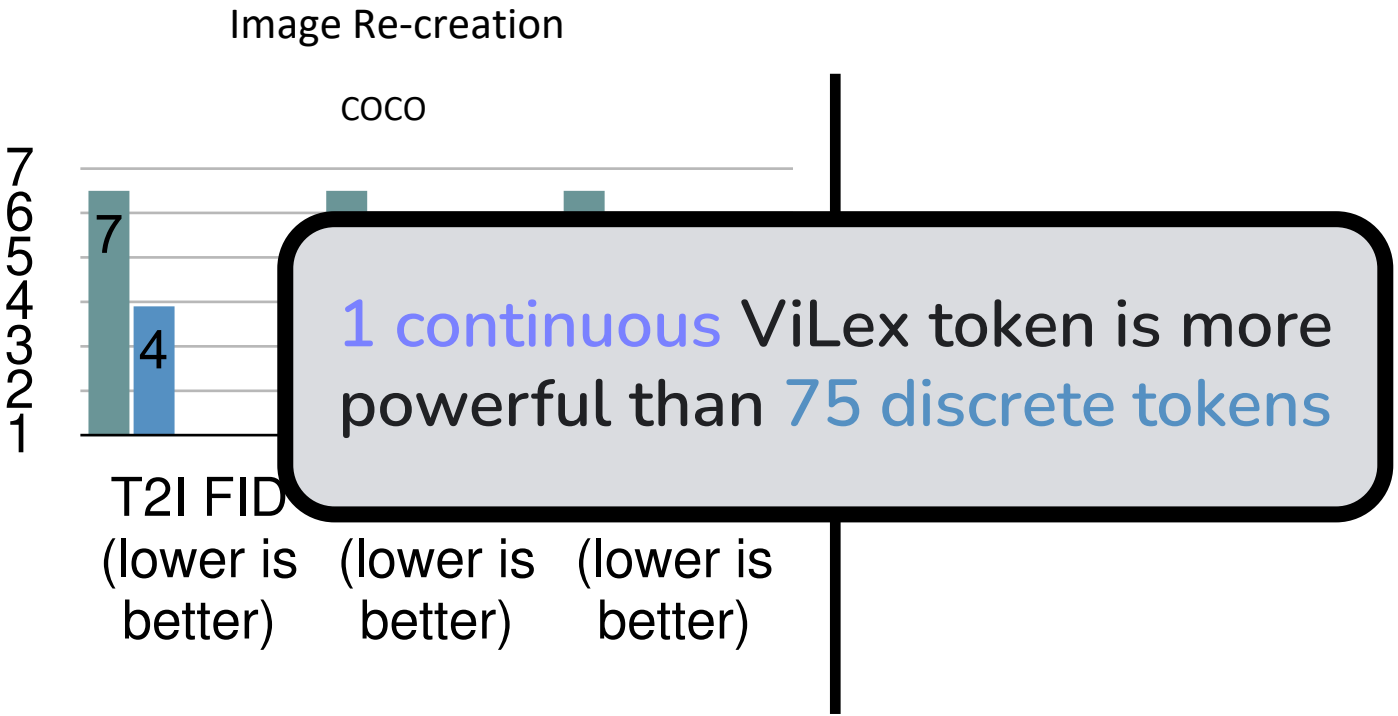
Generated Images
(Visual Prompt 1 + Text Prompt)

Generated Images
(Visual Prompt 2 + Text Prompt)

Generated Images
(Visual Prompt 3 + Text Prompt)

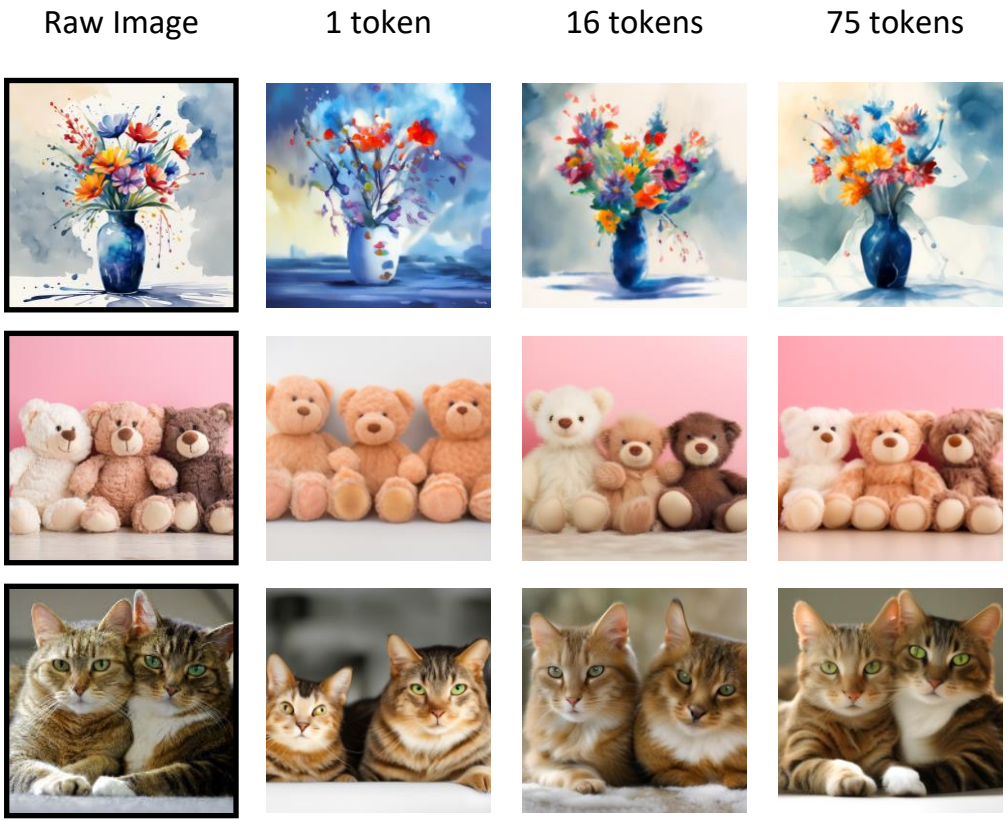
Wang, X., Zhou, X., Fathi, A., Darrell, T. and Schmid, C. "Visual Lexicon: Rich Image Features in Language Space" CVPR 2025.

ViLex: One Model for **BOTH** Understanding and Generation Tasks!



Wang et al. “Visual Lexicon: Rich Image Features in Language Space” CVPR 2025
Betker, James, et al. "Improving image generation with better captions." Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf 2.3 (2023): 8.
Wei, Chen, et al. "De-diffusion makes text a strong cross-modal interface." CVPR 2024

Which Visual Features Require More ViLex Tokens?

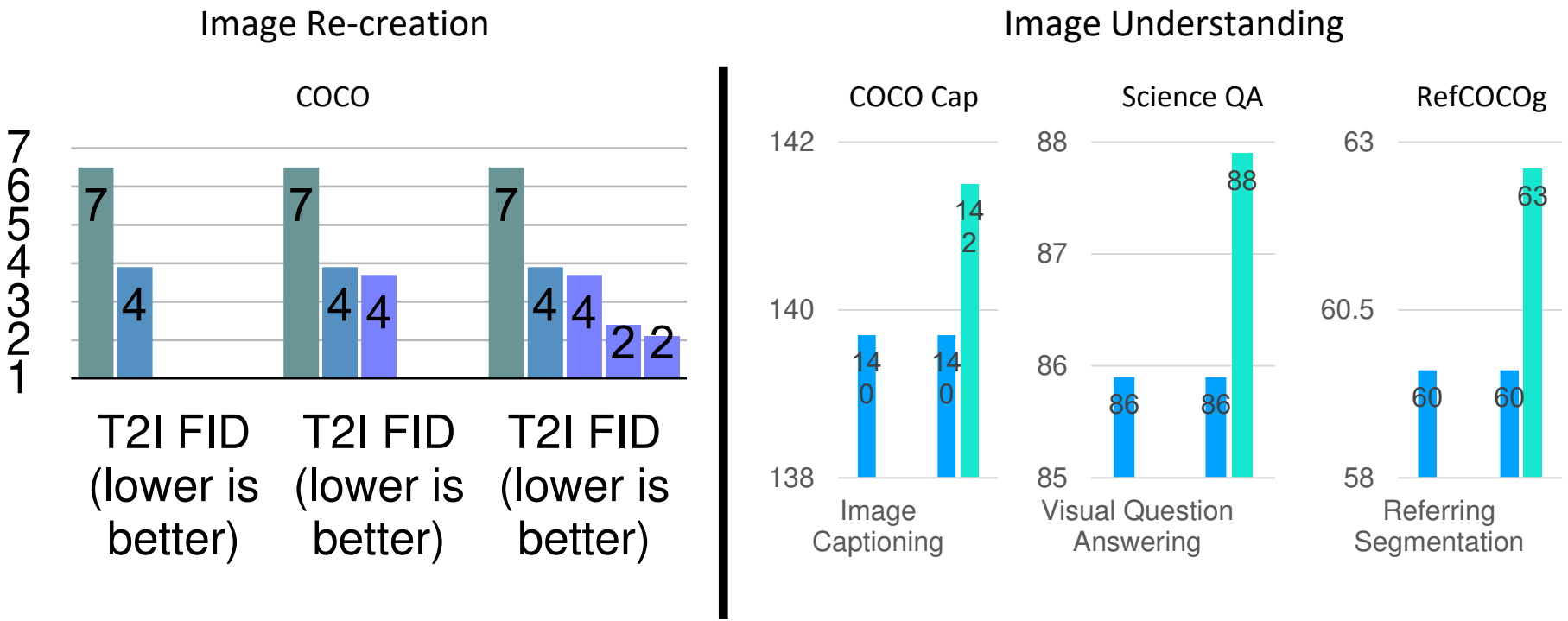


Early Tokens
Semantic classes, relative positions, layouts, etc

Late Tokens
Colors, textures, shapes, poses, etc

Wang, X., Zhou, X., Fathi, A., Darrell, T. and Schmid, C. “Visual Lexicon: Rich Image Features in Language Space” CVPR 2025.

ViLex: One Model for **BOTH** Understanding and Generation Tasks!



Wang et al. “Visual Lexicon: Rich Image Features in Language Space” CVPR 2025
Betker, James, et al. "Improving image generation with better captions." Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf 2.3 (2023): 8.
Wei, Chen, et al. "De-diffusion makes text a strong cross-modal interface." CVPR 2024

LLMs from Text to Vision and Robotics and back…

- Are LLMs Grounded? … *Surprisingly so!*
- Reducing VLM Hallucination with REVERSE Retrospective Sampling
- Efficient Scaling of VLMs to 4K Resolution with via PS3
- **Visual Tokens for Non-linguistic Generation (ViLex)**
- Navigation World Models
- 4D Reconstruction for Humanoid Robotics

LLMs from Text to Vision and Robotics and back…

- Are LLMs Grounded? … *Surprisingly so!*
- Reducing VLM Hallucination with REVERSE Retrospective Sampling
- Efficient Scaling of VLMs to 4K Resolution with via PS3
- Visual Tokens for Non-linguistic Generation (ViLex)
- **Navigation World Models**
- 4D Reconstruction for Humanoid Robotics

Navigation World Models

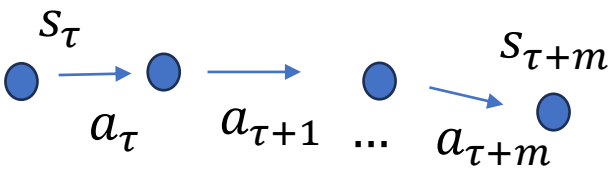
Lead author: Amir Bar
UC Berkeley, Meta

Construct a Predictive World Model primarily from visual inputs

Planning using Generative World Models

$s_{\tau+1} \sim f(s_\tau, a_\tau)$ f maps the current state and action to the future state.

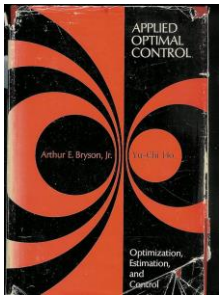
If we have a capable model of this form, we can use it to plan.



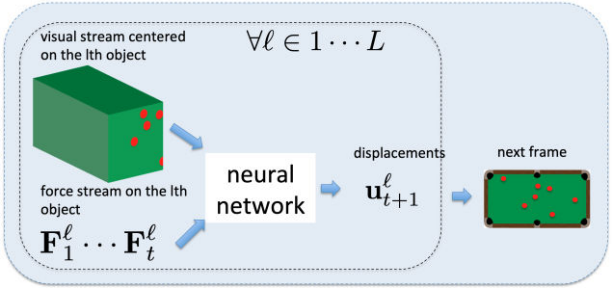
Evaluate

s^*

Action Conditioned Video Models as World Models



[Bryson and Ho, 1969]

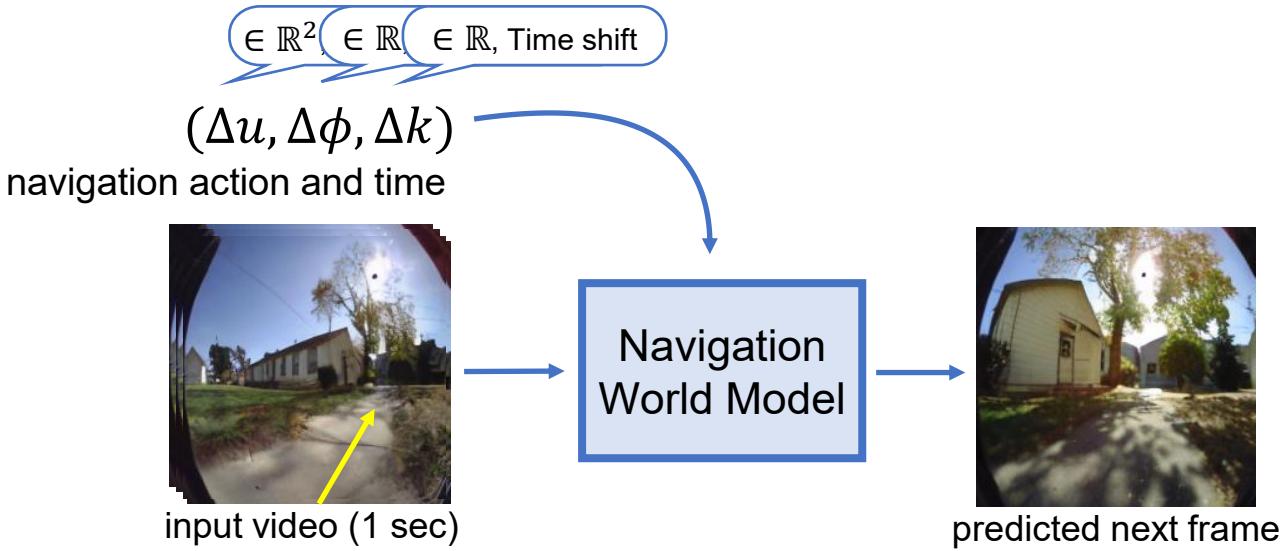


[Fragkiadaki et. al, 2015]

[Hafner et. al, 2023]

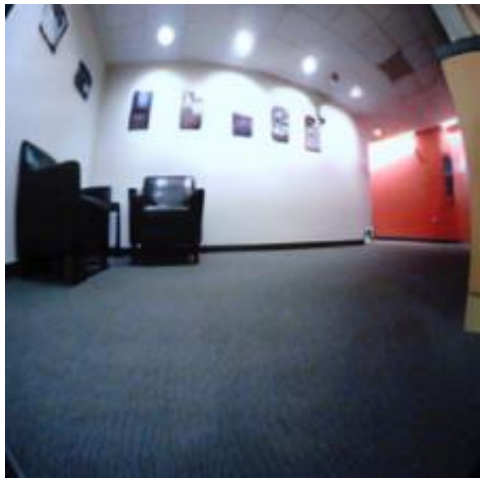
Goal: Trained from *offline video data*, *across* embodiments and environments, *without* rewards

Case study: Visual Navigation



A Bar, G Zhou, D Tran, T Darrell, Y LeCun, “Navigation World Models”. CVPR 2025.

Start



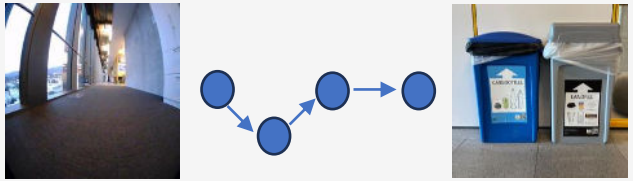
Goal Image

* **Video is generated** given first frame and actions

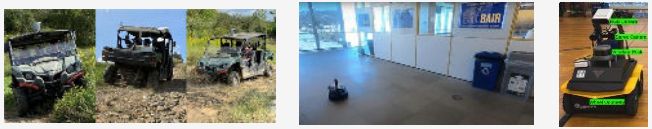
A Bar, G Zhou, D Tran, T Darrell, Y LeCun, “Navigation World Models”. CVPR 2025.

Navigation World Models

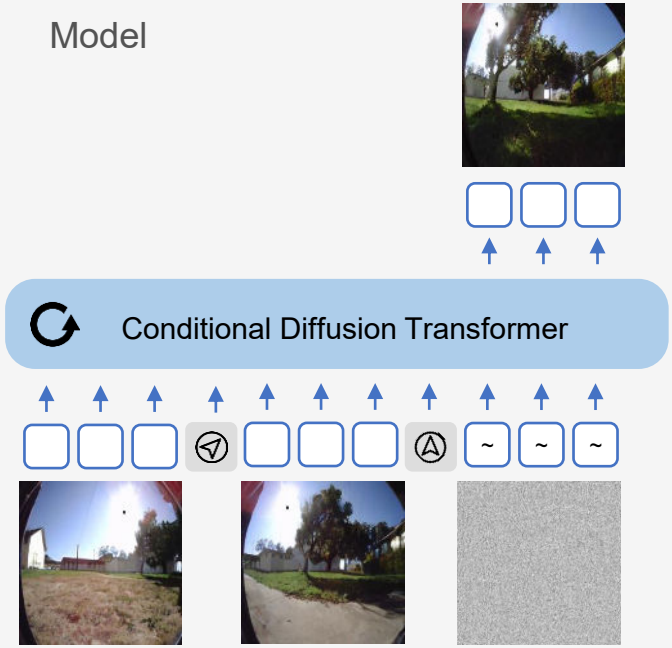
Planning



Pretraining data



Model



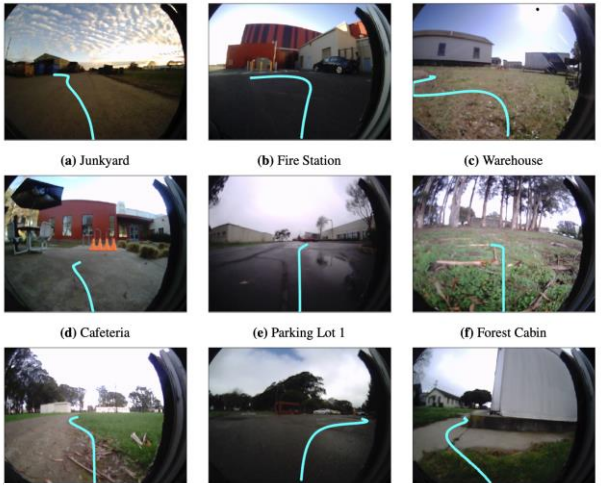
A Bar, G Zhou, D Tran, T Darrell, Y LeCun, “Navigation World Models”. CVPR 2025.

Cross Embodiment and Environment Data

SCAND [Xiao et. al, 2022]

TartanDrive [Triest et. al, 2022]

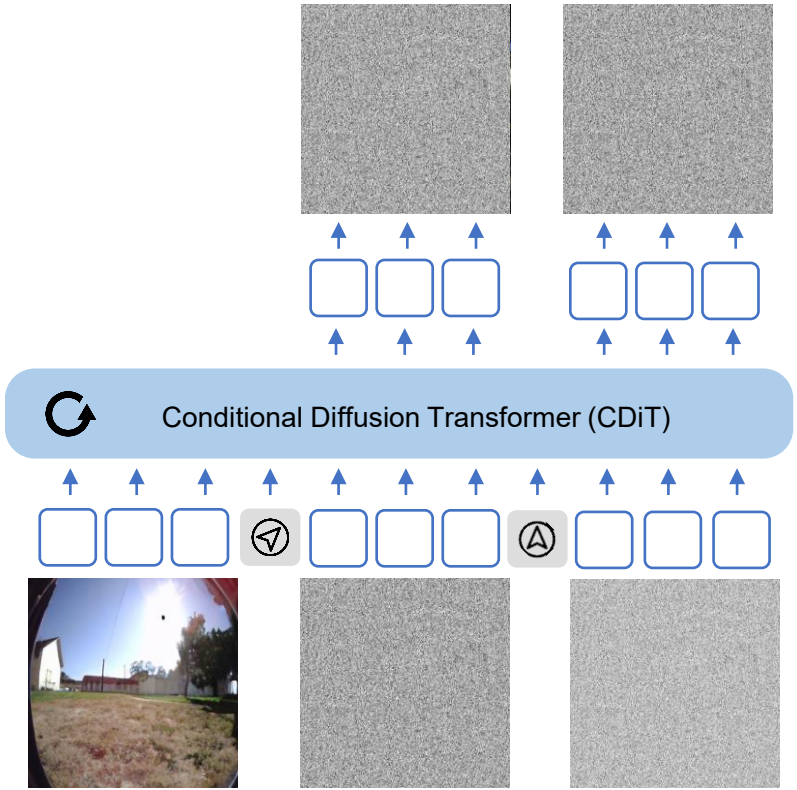
HuRon [Hirose et. al, 2023]



RECON [Shah et. al, 2021]

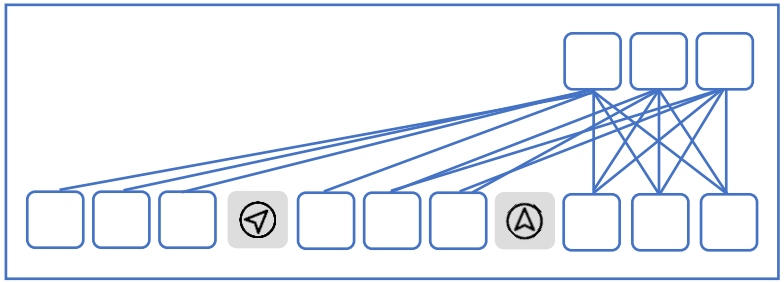
Videos and associated navigation actions
Diverse *environments* and *embodiments*

Conditional Diffusion Transformer



Attention in CDiT is **linear** in #frames

$$O(dmn^2 + nmd^2)$$



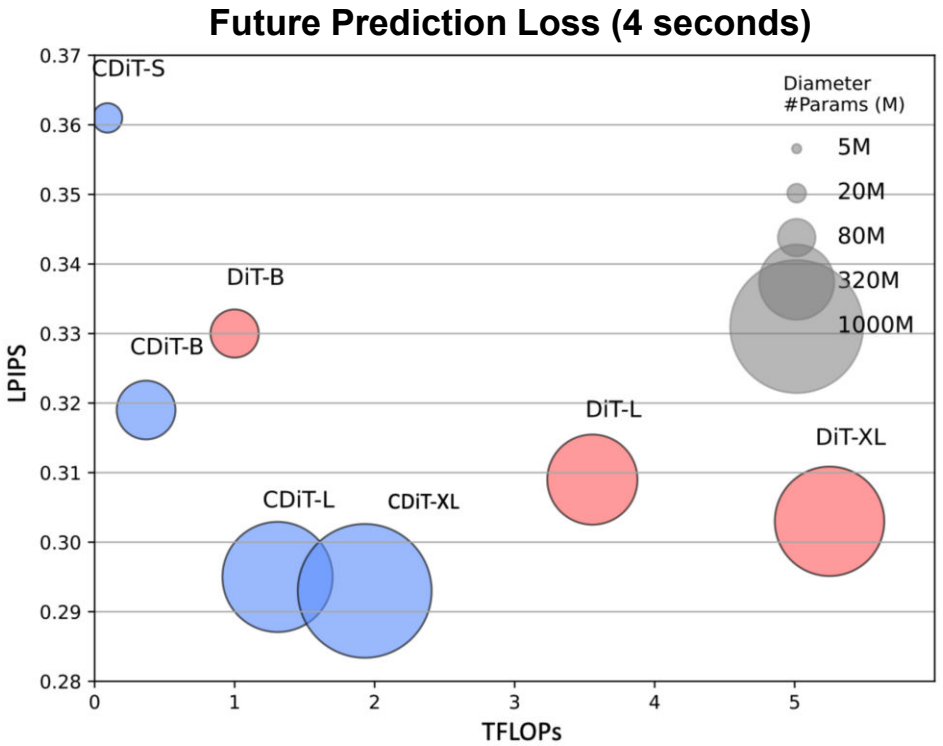
n - **#tokens per frame** m - **#frames** d - **#token dim**

A Bar, G Zhou, D Tran, T Darrell, Y LeCun, “Navigation World Models”. CVPR 2025.

Comparison to Diffusion Transformer

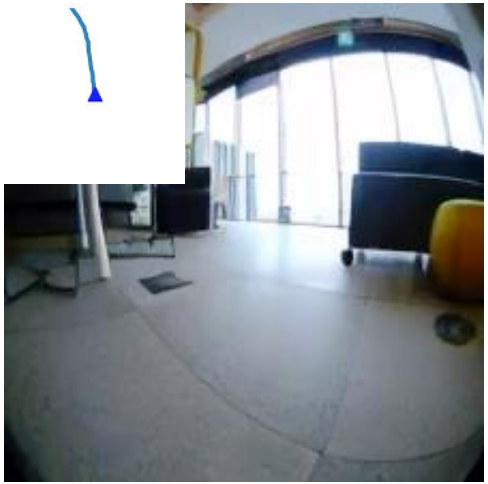
[Peebles et. al, 2023]

- Attention complexity (#frames)
  - DiT is Quadratic
  - CDiT is Linear
- 5x less Floating Point Operations



A Bar, G Zhou, D Tran, T Darrell, Y LeCun, “Navigation World Models”. CVPR 2025.

Follow Trajectories (*known environments*)



* **Video is generated** given first frame and actions

Follow Trajectories (*known environments*)

* **Video is generated** given first frame and actions

Follow Trajectories (*known environments*)

* **Video is generated** given first frame and actions

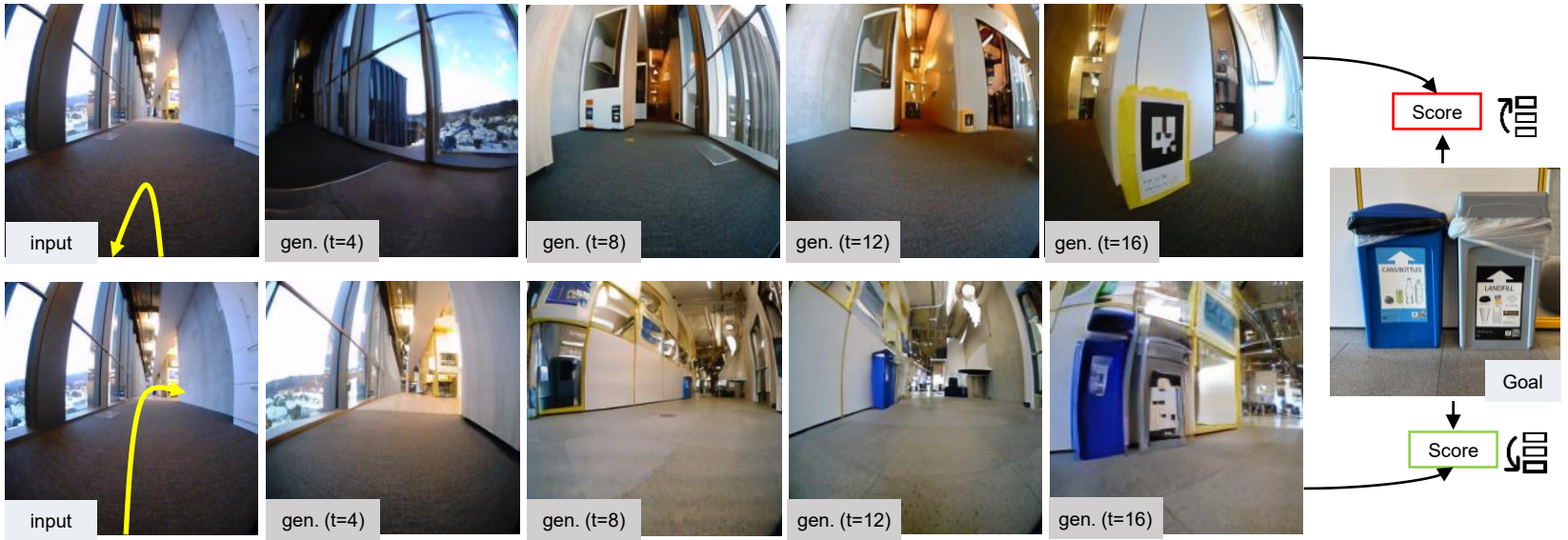
Follow Trajectories (un*known environments*)

* **Video is generated** given first frame and actions

Follow Trajectories (un*known environments*)

* **Video is generated** given first frame and actions

Navigation Planning



A Bar, G Zhou, D Tran, T Darrell, Y LeCun, “Navigation World Models”. CVPR 2025.

Planning with World Models

States Goal Actions

$$J(s_{\tau:\tau+m}, s^*, a_{\tau:\tau+m-1}) = \text{Distance to Goal} + \text{Action Constraints} + \text{State Constraints}$$

$$a_{\tau:\tau+m-1} = \arg\min_{a_{\tau:\tau+m-1}} \mathbb{E}_{\mathbf{s}}[J(s_{\tau:\tau+m}, s^*, a_{\tau:\tau+m-1})]$$

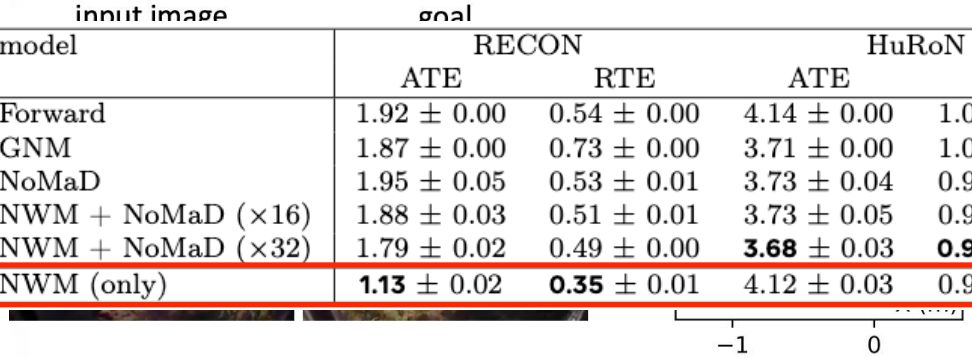
A Model Predictive Control (MPC) problem [Bryson and Ho, 1969].

Minimize the cost function using the Cross Entropy Method [Rubinstein, 1969].

A Bar, G Zhou, D Tran, T Darrell, Y LeCun, “Navigation World Models”. CVPR 2025.

Sample, Simulate and Rank

Sample trajectories from an external policy, like NoMaD [Sridhar et. al, 2024]



model	RECON		HuRoN		Tartan		SCAND	
	ATE	RTE	ATE	RTE	ATE	RTE	ATE	RTE
Forward	1.92 ± 0.00	0.54 ± 0.00	4.14 ± 0.00	1.05 ± 0.00	5.75 ± 0.00	1.19 ± 0.00	2.97 ± 0.00	0.62 ± 0.00
GNM	1.87 ± 0.00	0.73 ± 0.00	3.71 ± 0.00	1.00 ± 0.00	6.65 ± 0.00	1.62 ± 0.00	2.12 ± 0.00	0.61 ± 0.00
NoMaD	1.95 ± 0.05	0.53 ± 0.01	3.73 ± 0.04	0.96 ± 0.01	6.32 ± 0.03	1.31 ± 0.01	2.24 ± 0.03	0.49 ± 0.01
NWM + NoMaD (×16)	1.88 ± 0.03	0.51 ± 0.01	3.73 ± 0.05	0.95 ± 0.01	6.26 ± 0.06	1.30 ± 0.01	2.18 ± 0.05	0.48 ± 0.01
NWM + NoMaD (×32)	1.79 ± 0.02	0.49 ± 0.00	**3.68** ± 0.03	**0.95** ± 0.01	6.25 ± 0.05	1.29 ± 0.01	2.19 ± 0.03	0.47 ± 0.01
NWM (only)	**1.13** ± 0.02	**0.35** ± 0.01	4.12 ± 0.03	0.96 ± 0.01	**5.63** ± 0.06	**1.18** ± 0.01	**1.28** ± 0.02	**0.33** ± 0.01

A Bar, G Zhou, D Tran, T Darrell, Y LeCun, “Navigation World Models”. CVPR 2025.

Navigation World Models - Takeaways

- Model dynamics using *Generative World Models*
- Plan by search, dynamically allocate compute
- More general than training a policy

A Bar, G Zhou, D Tran, T Darrell, Y LeCun, “Navigation World Models”. CVPR 2025.

LLMs from Text to Vision and Robotics and back…

- Are LLMs Grounded? … *Surprisingly so!*
- Reducing VLM Hallucination with REVERSE Retrospective Sampling
- Efficient Scaling of VLMs to 4K Resolution with via PS3
- Visual Tokens for Non-linguistic Generation (ViLex)
- **Navigation World Models with CDiT**
- 4D Reconstruction for Humanoid Robotics

LLMs from Text to Vision and Robotics and back…

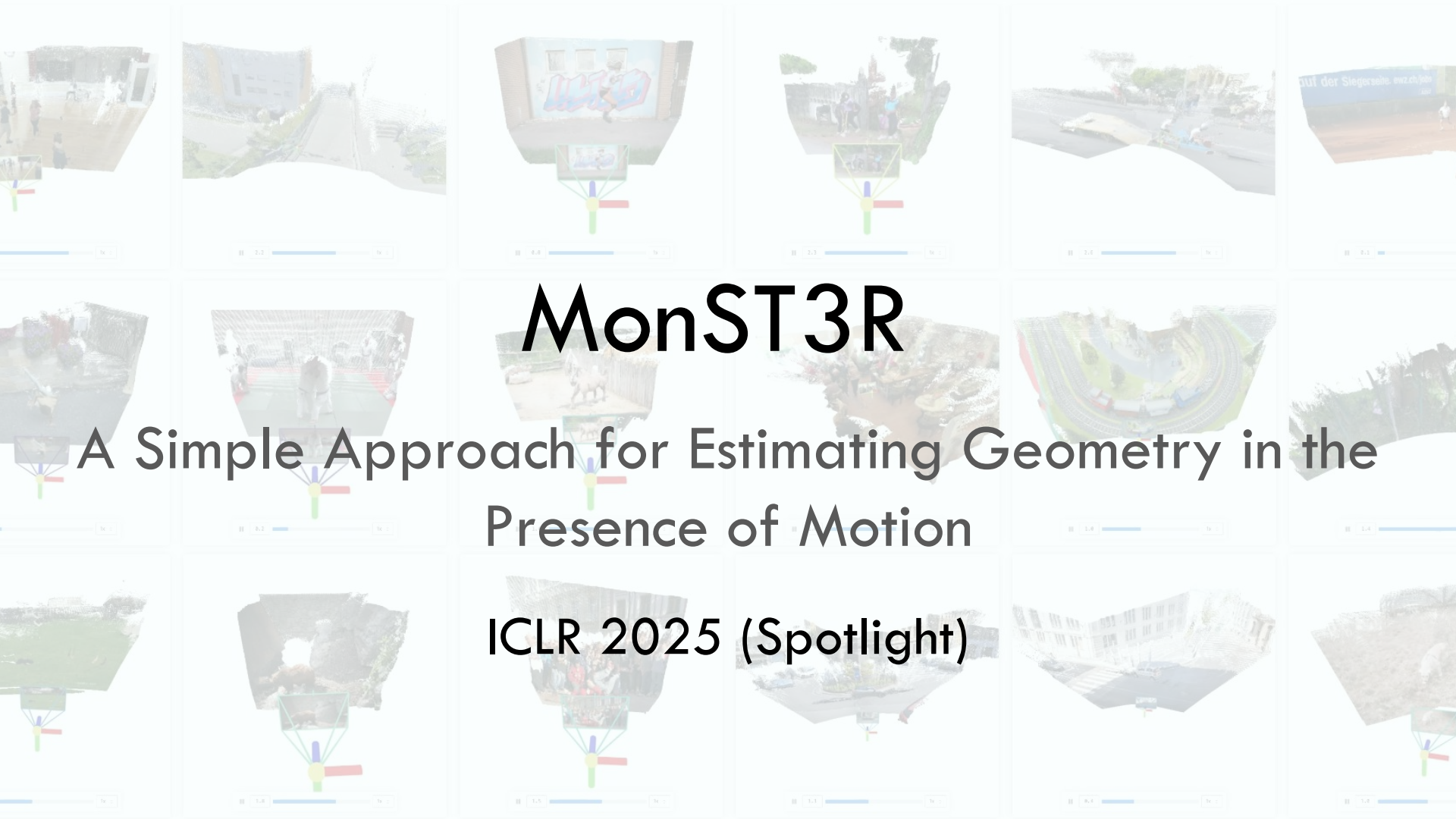
- Are LLMs Grounded? … *Surprisingly so!*
- Reducing VLM Hallucination with REVERSE Retrospective Sampling
- Efficient Scaling of VLMs to 4K Resolution with via PS3
- Visual Tokens for Non-linguistic Generation (ViLex)
- Navigation World Models with CDiT
- **4D Reconstruction for Humanoid Robotics**

A Simple Approach to Estimating Geometry and Motion from Casual Videos

Lead Author: Junyi Zhang

How to Efficiently Estimate **Geometry** and **Motion** from Casual Videos

How to Apply these Techniques to Robotics



MonST3R

A Simple Approach for Estimating Geometry in the Presence of Motion

ICLR 2025 (Spotlight)

BERKELEY ARTIFICIAL INTELLIGENCE RESEARCH

Junyi Zhang

Charles Herrmann+

Junhwa Hur

Varun Jampani

Trevor Darrell

Forrester Cole

Deqing Sun*

Ming-Hsuan Yang*

Overview

Given an unposed video of dynamic scene

Overview

Video Input

We reconstruct dynamic point cloud & camera poses

Overview

Video Input

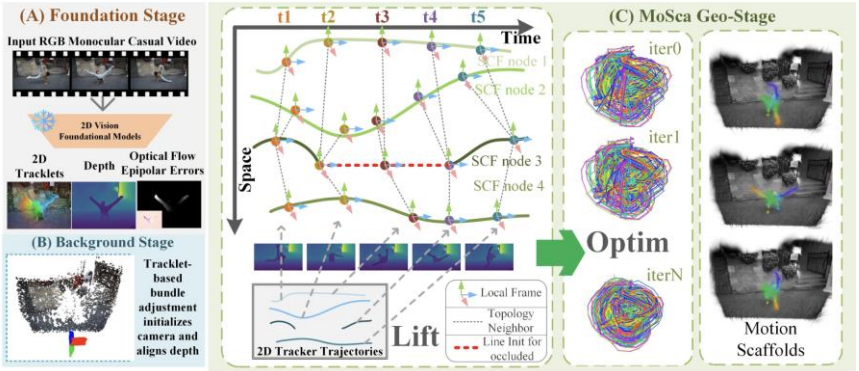
We can also pull-out video depth & motion mask

Background – Dynamic 3D reconstruction

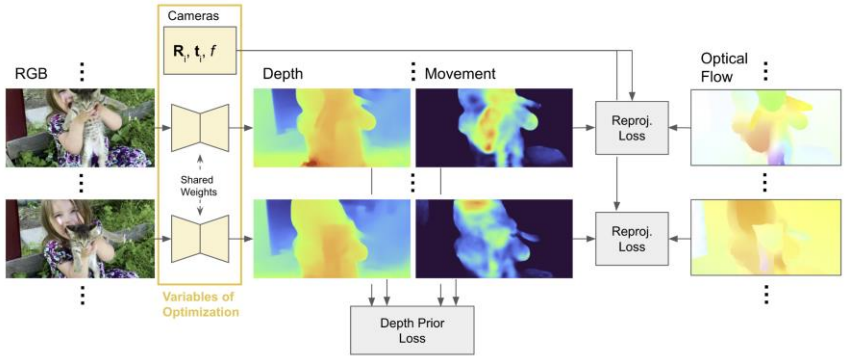
Dynamic 3D reconstruction is an important problem for:

1. autonomous driving, 2. robotics, 3. AR/VR, 4. spatial understanding…

Existing methods target it as separate problems and use heavy optimization



MoSca [1]



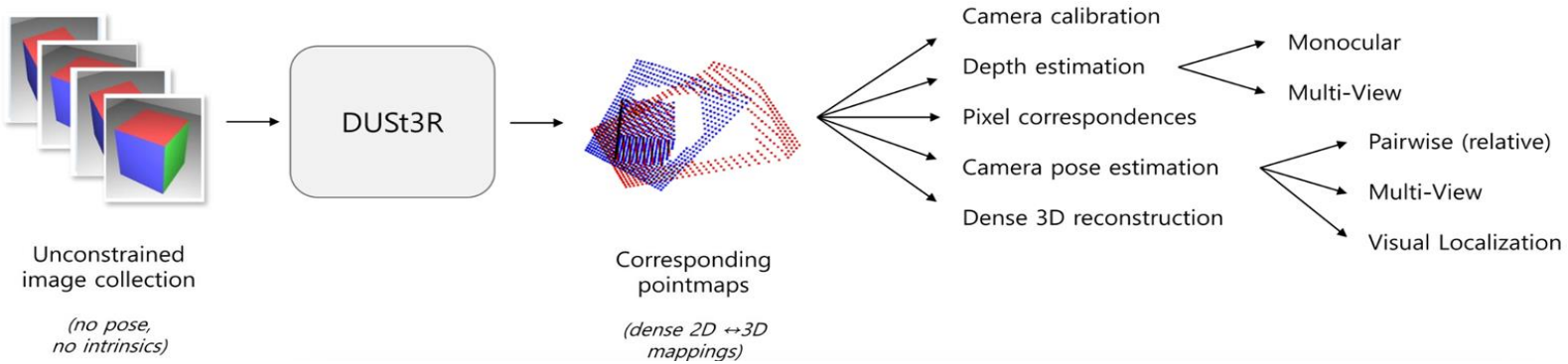
CasualSAM [2]

[1] Lei Jiahui, et al. MoSca: Dynamic Gaussian Fusion from Casual Videos via 4D Motion Scaffolds. *Arxiv 2024*

[2] Zhang Zhoutong, et al. Structure and Motion from Casual Videos. *ECCV 2022*

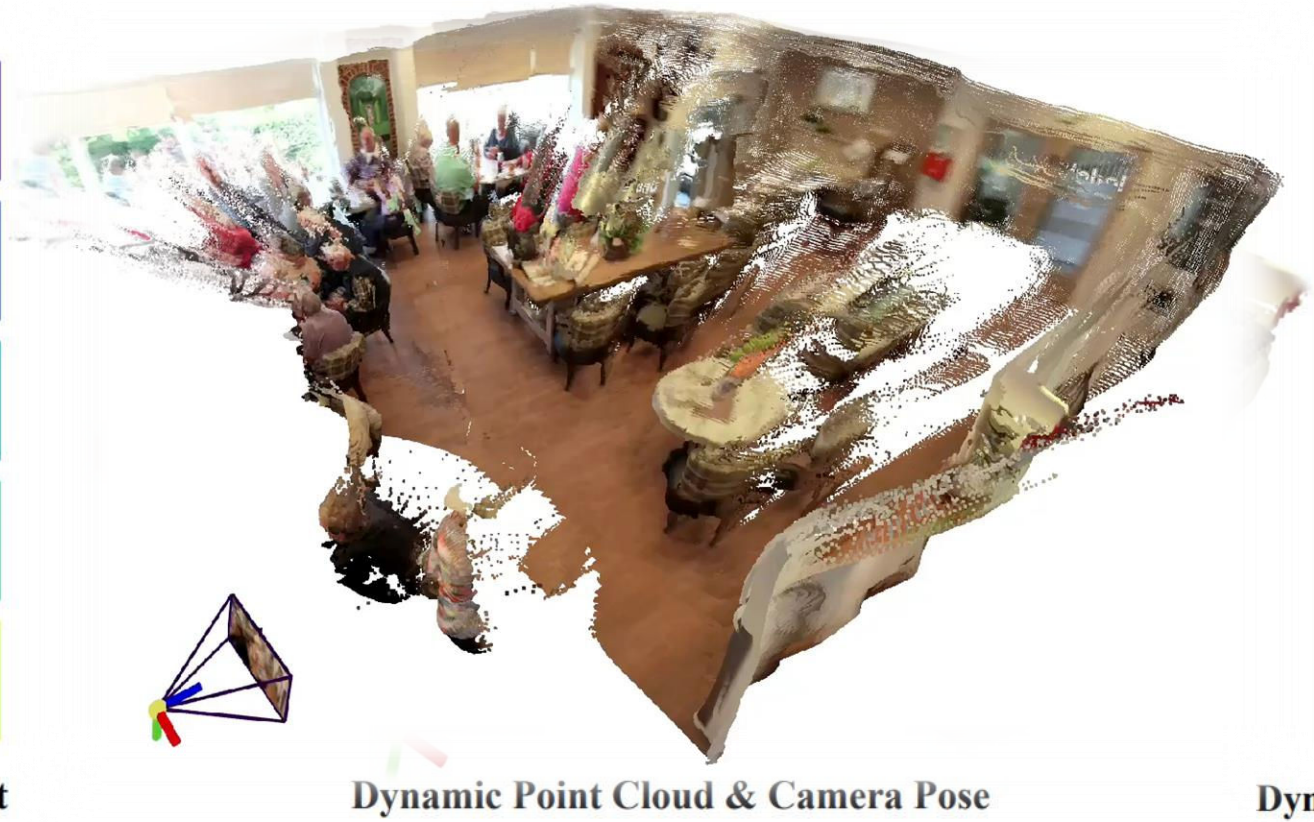
Background – DUSt3R

DUSt3R seeks to unify and simplify all 3D vision tasks with the pointmap representations



Can we also target dynamics in a feed-forward way? This would be a superset of DUSt3R since 3D is just "video" without dynamics

Video Input



Dynamic Point Cloud & Camera Pose

Video Depth

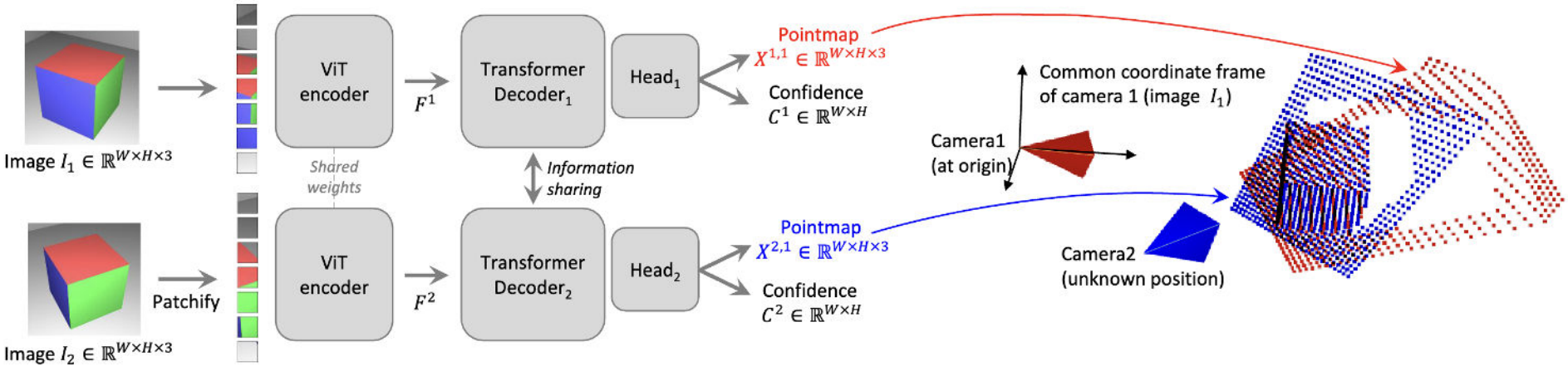
Camera Intrinsics

Dynamic / Static Mask

Key Insight

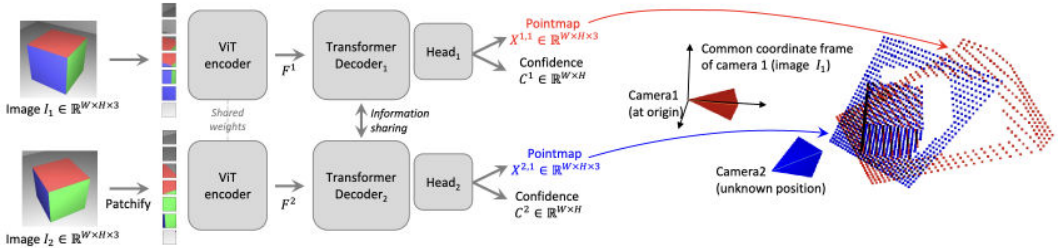
Pointmap representation

Estimate xyz coordinates for two frames, aligned in the camera coordinate system of frame1 as they would be in the real world.

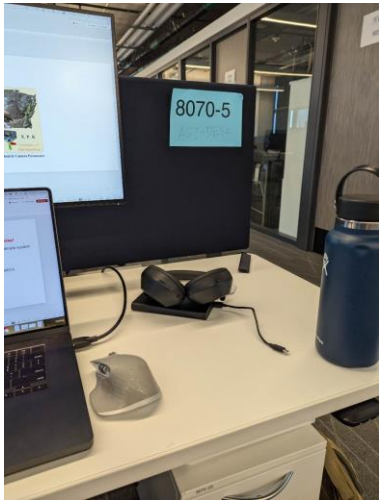


Key Insight

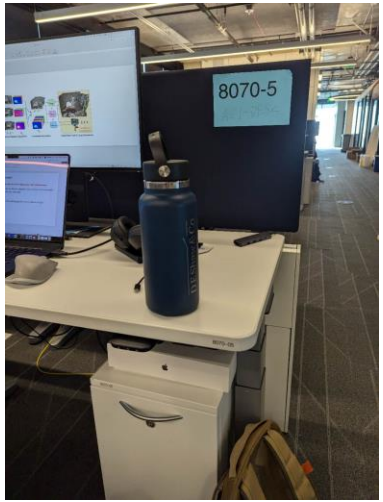
Pointmap representation



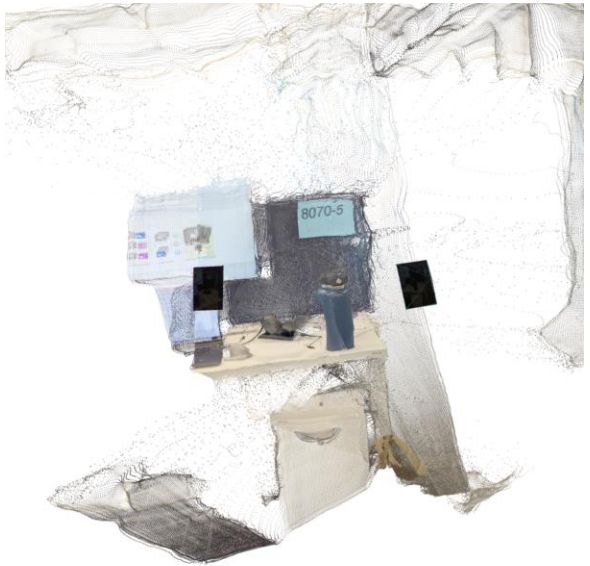
Estimate xyz coordinates for two frames, aligned in the camera coordinate system of frame1 as they would be in the real world.



Frame 1



Frame 2



Point cloud

"**Multi-view** of the same **static scene**"

What’s missing in MonST3R?

Getting rid of global alignment!

Correspondence!

Can we have them all?

And with the same architecture!

St4RTrack

input

overlap

Simultaneous 4D Reconstruction and Tracking in the World

tracking

reconstruction

Junyi Zhang*

Haiwen Feng*

Qianqian Wang

Yufei Ye

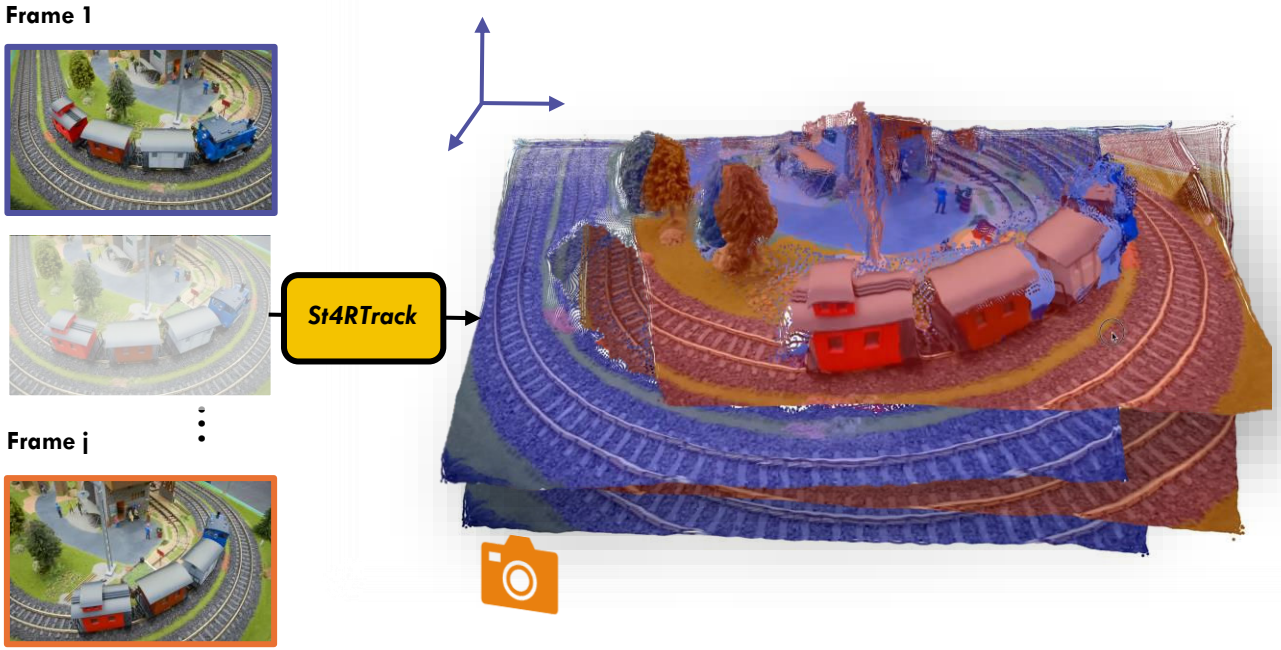
Pengcheng Yu

Michale J. Black

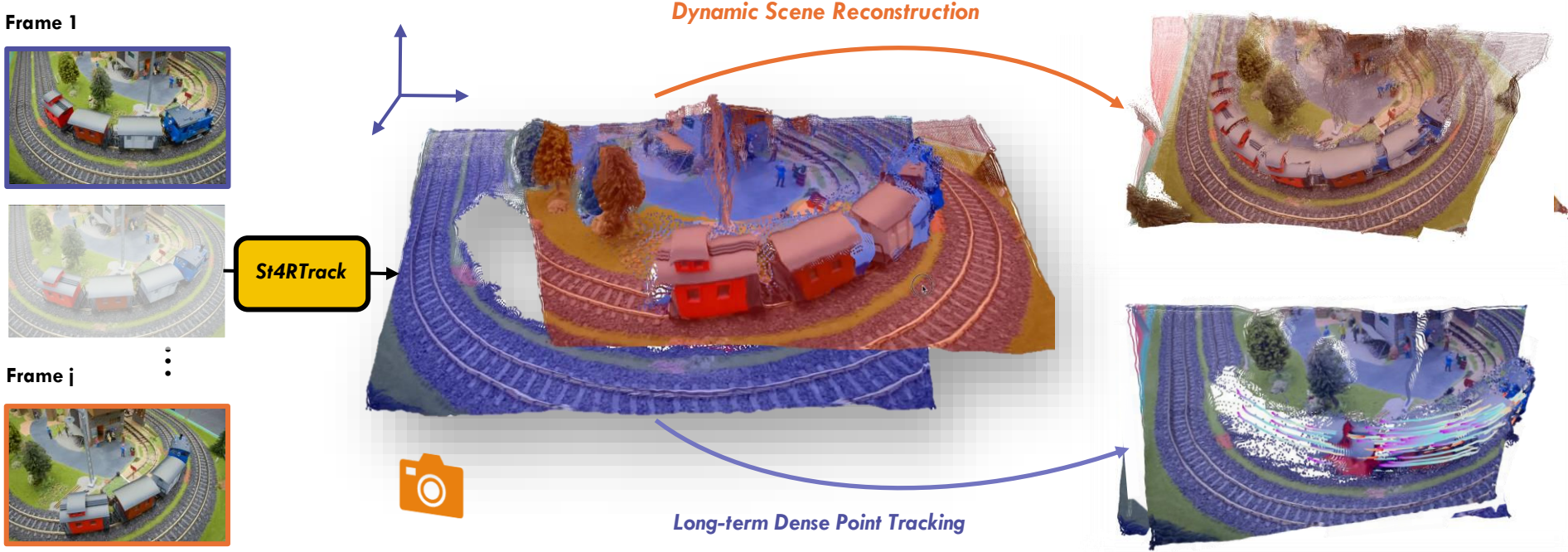
Trevor Darrell

Angjoo Kanazawa

Overview

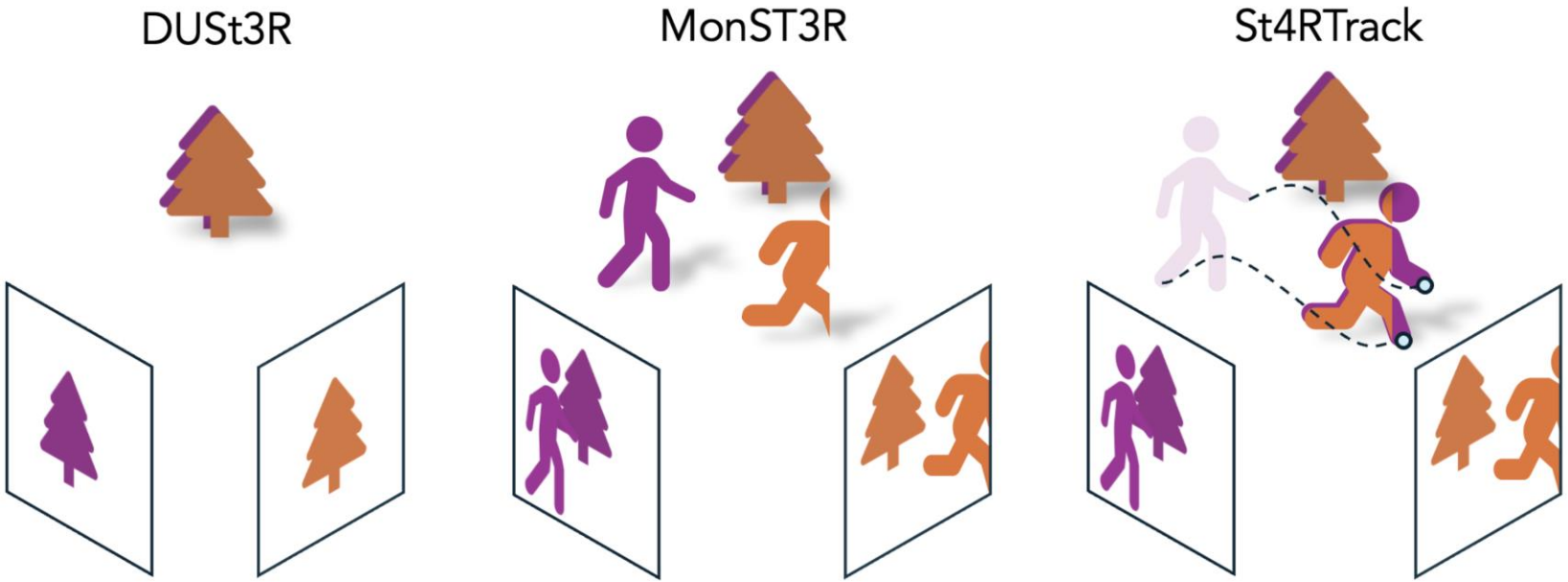


Overview


Frame 1
Frame j
St4RTrack
Dynamic Scene Reconstruction
Long-term Dense Point Tracking

Unified 4D Modeling with Time-Dependent Pointmap

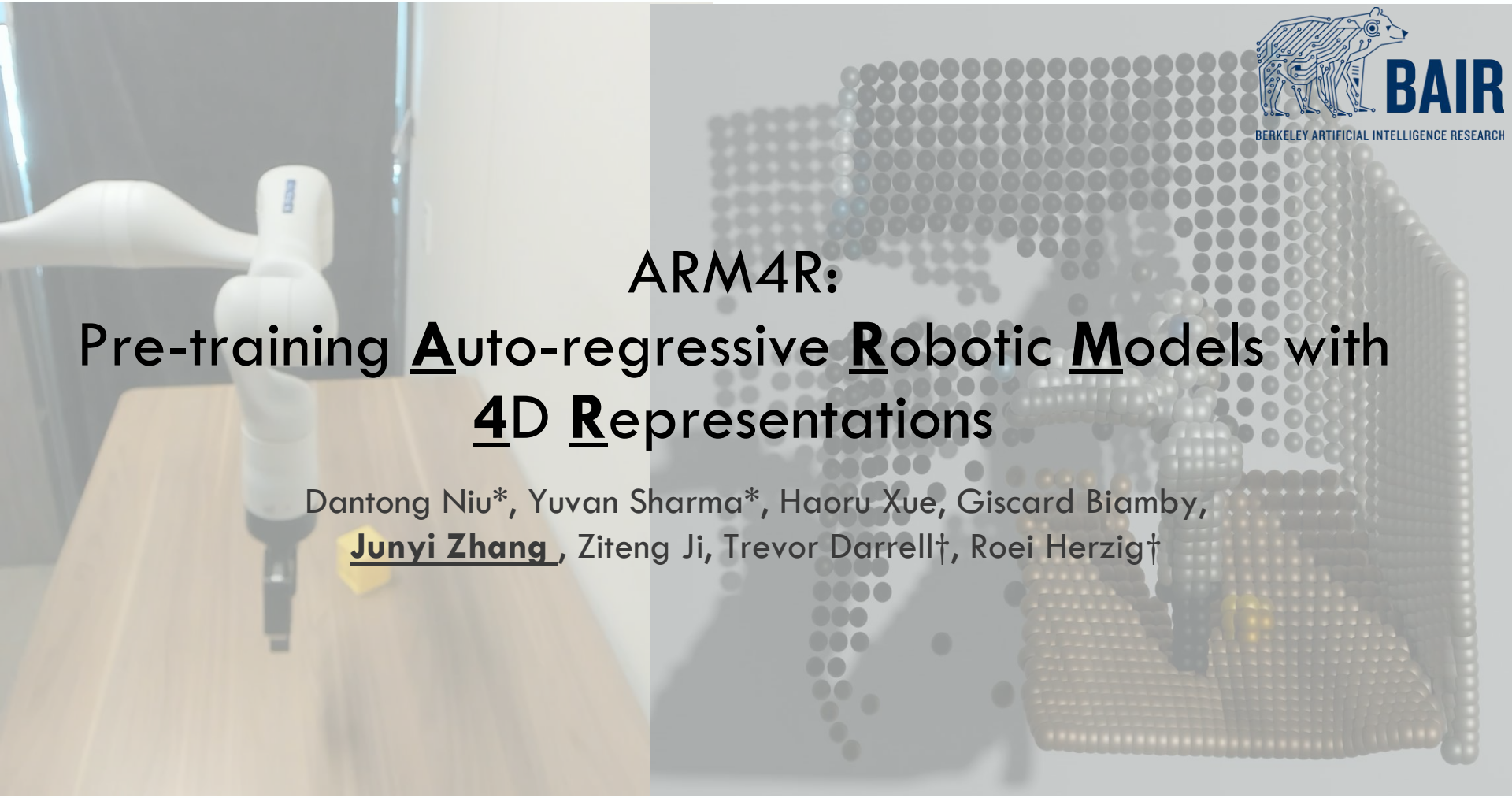
Core Idea: by properly defining the two pointmaps, we can enable simultaneous tracking and reconstruction, using the same architecture!



Conceptual difference between DUSt3R, MonST3R, and St4RTrack

St4RTrack estimates two pointmaps at the same timestamp. It predicts how the **points in the first frame move** to the second frame, and **reconstructs the geometry** of the second frame.

What could 3D motion be used for?



ARM4R: Pre-training **A**uto-regressive **R**obotic **M**odels with **4**D **R**epresentations

Dantong Niu*, Yuvan Sharma*, Haoru Xue, Giscard Biamby,
Junyi Zhang, Ziteng Ji, Trevor Darrell†, Roei Herzig†

BAIR
BERKELEY ARTIFICIAL INTELLIGENCE RESEARCH

Motivation

Existing Pre-training work:

- MVP: Pretraining the vision encoder on ego-centric data
- RPT: Pretraining the policy transformer on visual signal + proprios
- LLARVA/OpenVLA: Utilizing a language decoder that has been pre-trained on semantic tasks such as VQA and image captioning.

4D representation is more spatially grounded and generalizable for robotic model pretraining

Approach

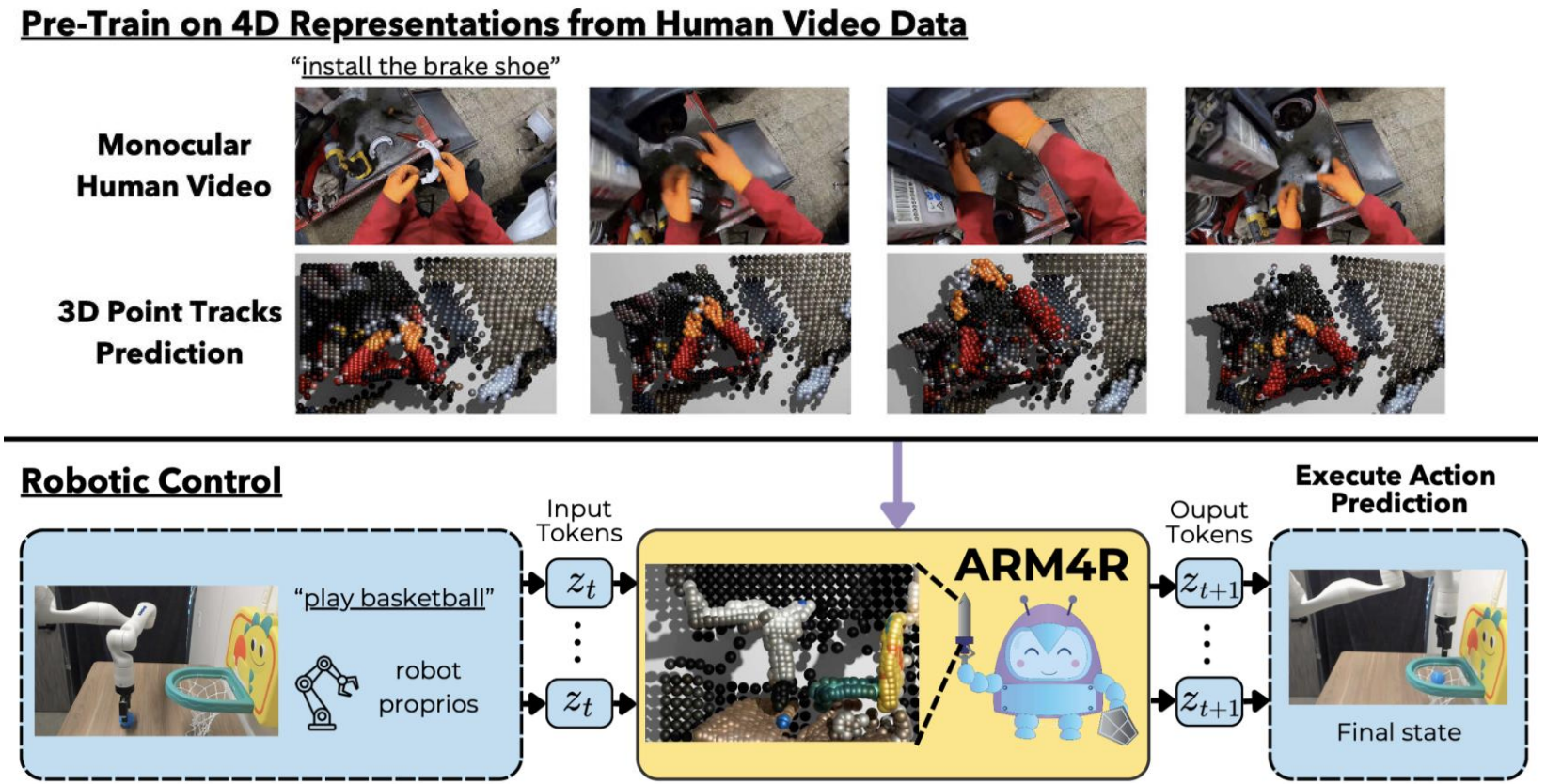
Pre-Train on 4D Representations from Human Video Data

"install the brake shoe"

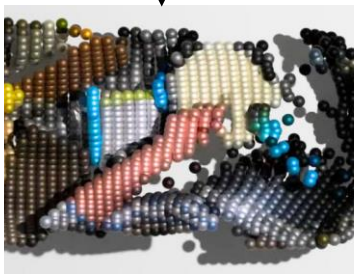
Monocular Human Video

3D Point Tracks Prediction

Approach


Pre-Train on 4D Representations from Human Video Data
"install the brake shoe"
Monocular Human Video
3D Point Tracks Prediction
Robotic Control
"play basketball"
robot proprios
Input Tokens
z_t
z_t
ARM4R
Ouput Tokens
z_{t+1}
z_{t+1}
Execute Action Prediction
Final state

Architecture



$\hat{p}_{t+1}$

Pre-Training: 3D Point Track Prediction

Decoder

z_l z_t^{obs} $\hat{z}_t$ • • •

Causal Transformer

• • • z_l z_t^{obs} $\hat{z}_t$ • • •

z_t^{im} Attention Pooling z_t^s

Language Encoder | Image Encoder | State Encoder | Future State Encoder

Fine-Tuning: Robot Control

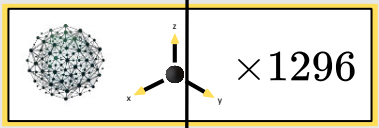
$\hat{s}_{t+1}$

3D Point Track Pre-Training and Fine-Tuning Stages

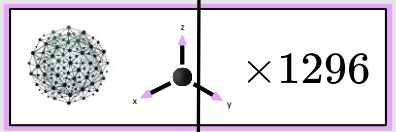
check temperature

Language Input l

Image Input i_t



Current Point Coordinates p_t



Future Point Coordinates p_{t+1}

Robot Control Fine-tuning Stage

play basketball

Language Input l

Image Input i_t

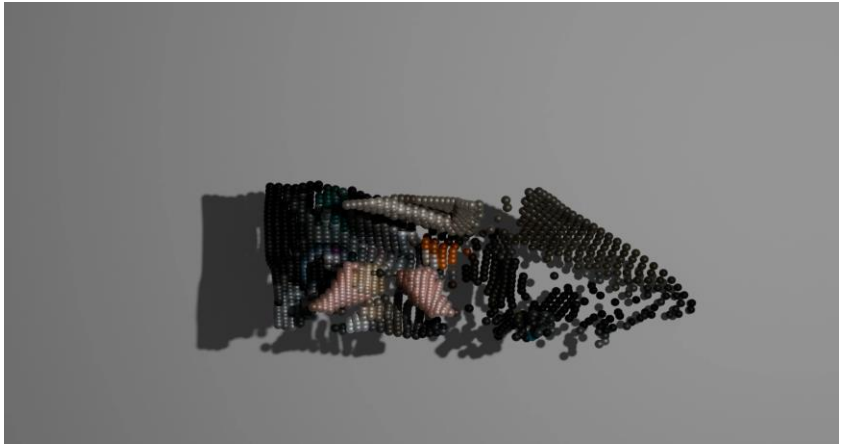
Robot Proprios

Current Robot State s_t

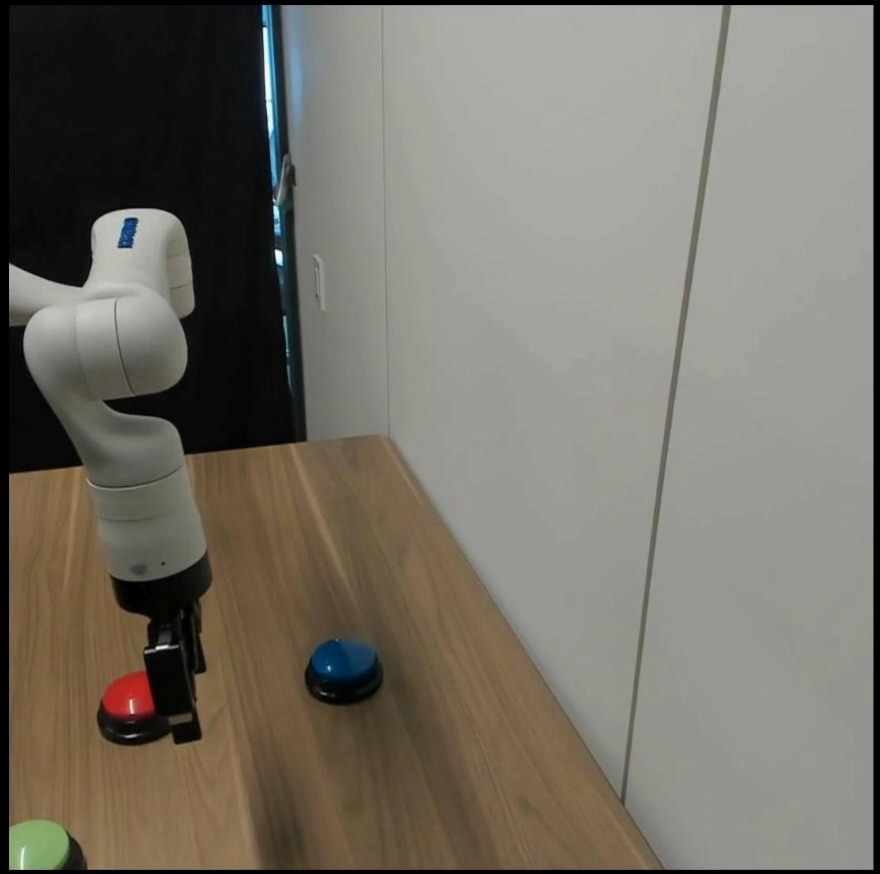
Future Robot Proprios

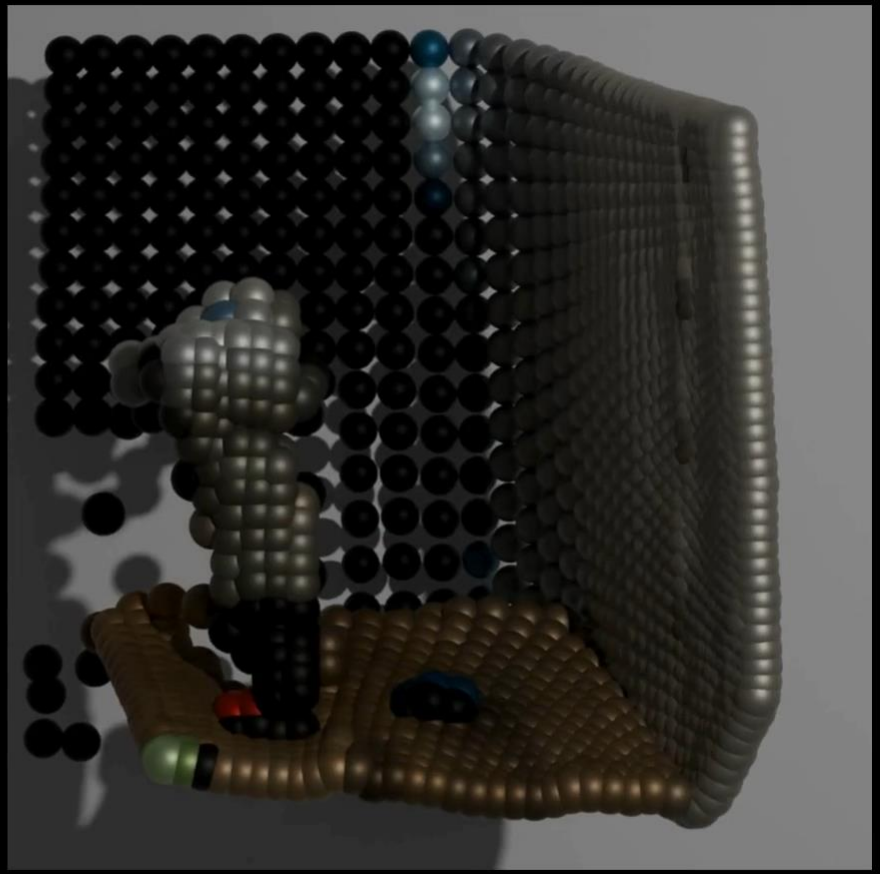
Future Robot State s_{t+1}

Pretraining results



“push red button”





Real-world Results

Comparison with baselines. Success rate (%) on the real Kinova Multi-Task setting

Method	pick cube up			destack		stack	
	yellow	cyan	green	yellow	cyan	yellow on cyan	cyan on yellow
ATM	5.3 ± 3.5	6.7 ± 2.7	9.3 ± 1.3	4.0 ± 2.3	9.3 ± 3.5	1.3 ± 1.3	2.6 ± 1.3
OpenVLA	77.8 ± 6.4	45.8 ± 4.2	91.7 ± 8.3	55.6 ± 2.8	51.3 ± 2.6	27.8 ± 2.8	38.5 ± 4.4
Ours	$\mathbf{92.6 \pm 3.7}$	$\mathbf{100 \pm 0.0}$	$\mathbf{95.8 \pm 4.2}$	$\mathbf{94.4 \pm 2.7}$	$\mathbf{94.9 \pm 5.1}$	$\mathbf{63.6 \pm 5.2}$	$\mathbf{59.5 \pm 2.4}$

Method	pick toys then place to target				push		Average
	spiderman	penguin	pig	play basketball	push red button	push red the blue	
ATM	5.3 ± 1.3	6.7 ± 1.3	5.3 ± 3.5	24.0 ± 4.6	4.0 ± 2.3	0.0 ± 0.0	6.4 ± 2.2
OpenVLA	2.7 ± 1.3	17.3 ± 1.3	2.7 ± 2.7	49.3 ± 3.5	23.1 ± 4.4	0.0 ± 0.0	37.2 ± 3.4
Ours	$\mathbf{90.7 \pm 1.3}$	$\mathbf{94.7 \pm 1.3}$	$\mathbf{93.3 \pm 1.3}$	$\mathbf{92.0 \pm 2.3}$	$\mathbf{84.6 \pm 4.4}$	$\mathbf{25.0 \pm 4.8}$	$\mathbf{83.1 \pm 3.0}$

Pre-training approaches comparison.
We compare to several others on pre-training on three tasks with a Kinova robot.

Method	pick cube	stack cubes	destack cubes
MVP	75.00	18.75	81.25
RPT	87.50	31.25	93.75
Octo	56.25	12.50	37.50
ATM	7.11	2.00	6.67
OpenVLA	68.75	31.25	53.33
LLARVA	93.75	56.25	**100.00**
ARM4R	$\mathbf{96.0 \pm 2.3}$	$\mathbf{61.3 \pm 1.3}$	94.7 ± 1.3

Sim Results

- Success rate (%) on the real RLBench Multi-Task setting.
- Comparing ARM4R's performance against several related baselines on 12 tasks from the RLBench benchmark

Method	Task												
	open drawer	meat off grill	turn tap	put money	push buttons	sweep dustpan	slide block	close jar	screw bulb	place wine	reach and drag	stack blocks	Average Success Rate (%)
Image-BC (ViT)	0	0	16	0	0	0	0	0	16	0	0	0	2.67
C2FARM-BC	20	20	68	12	**72**	0	16	24	8	18	24	4	23.83
ManiGaussian	76	60	56	-	20	64	24	28	-	-	**92**	12	48.00
LLARVA	60	80	56	44	56	**84**	**100**	28	8	12	52	0	48.33
PerAct	80	84	**80**	44	48	56	72	**60**	**24**	12	68	**36**	55.33
ARM4R	**88.8**	**94.4**	61.6	**92.0**	67.2	72.0	85.6	24.0	10.4	**36.0**	77.6	4.0	**59.47**

What could 3D geometry be used for?

VideoMimic

Imitating Human Motion over Terrains from
Reconstructed Casually Captured Video

Arthur Allshire, Hongsuk Ben Choi, David McAllister, Junyi Zhang
(alphabetical order)

Motivation

Many great works in humanoid locomotion

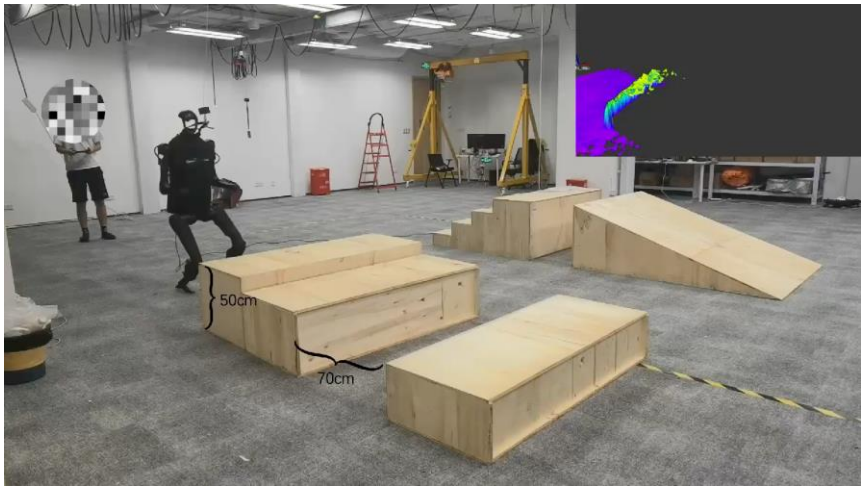
→ *locomotion, whole body control, parkour, etc.*

Radosavovic et al. 2024

Motivation

Many great works in humanoid locomotion

→ *locomotion, whole body control, parkour, etc.*



Long et al. 2024

Motivation

Unfortunately, specifying rewards to get great humanoid control has become **too complex**

Why? annotation for robots is expensive!

TABLE I: Rewards

Reward	Equation (r_i)	Weight (w_i)
Lin. velocity tracking	$\exp\left\{-\frac{\lVert \mathbf{v}_{xy}^{\text{cmd}} - \mathbf{v}_{xy} \rVert_2^2}{\sigma}\right\}$	1.0
Ang. velocity tracking	$\exp\left\{-\frac{\left(\omega_{\text{yaw}}^{\text{cmd}} - \omega_{\text{yaw}}\right)^2}{\sigma}\right\}$	1.0
Linear velocity (z)	v_z^2	-0.5
Angular velocity (xy)	$\lVert \boldsymbol{\omega}_{xy} \rVert_2^2$	-0.025
Orientation	$\lVert \mathbf{g}_x \rVert_2^2 + \lVert \mathbf{g}_y \rVert_2^2$	-1.25
Joint accelerations	$\lVert \ddot{\boldsymbol{\theta}} \rVert_2^2$	-2.5×10^{-7}
Joint power	$\frac{\lvert \boldsymbol{\tau} \lVert \dot{\boldsymbol{\theta}} \rvert^T}{\lVert \mathbf{v} \rVert_2^2 + 0.2 * \lVert \boldsymbol{\omega} \rVert_2^2}$	-2.5×10^{-5}
Body height w.r.t. feet	$\left(h^{\text{target}} - h\right)^2$	0.1
Feet clearance	$\sum_{feet} \left(p_z^{\text{target}} - p_z^i\right)^2 \cdot v_{xy}^i$	-0.25
Action rate	$\lVert \mathbf{a}_t - \mathbf{a}_{t-1} \rVert_2^2$	-0.01
Smoothness	$\lVert \mathbf{a}_t - 2\mathbf{a}_{t-1} + \mathbf{a}_{t-2} \rVert_2^2$	-0.01
Feet stumble	$\mathbf{1}\left\{\exists i, \lvert \mathbf{F}_i^{xy} \rvert > 3 \lvert F_i^z \rvert\right\}$	-3.0
Torques	$\sum_{\text{all joints}} \lvert \frac{\tau_i}{kp_i} \rvert_2^2$	-2.5×10^{-6}
Joint velocity	$\sum_{\text{all joints}} \lvert \dot{\theta}_i \rvert_2^2$	-1×10^{-4}
Joint tracking error	$\sum_{\text{all joints}} \lvert \theta_i - \theta_i^{target} \rvert^2$	-0.25
Arm joint deviation	$\sum_{\text{arm joints}} \lvert \theta_i - \theta_i^{default} \rvert^2$	-0.1
Hip joint deviation	$\sum_{\text{hip joints}} \lvert \theta_i - \theta_i^{default} \rvert^2$	-0.5
Waist joint deviation	$\sum_{\text{waist joints}} \lvert \theta_i - \theta_i^{default} \rvert^2$	-0.25
Joint pos limits	$\sum_{\text{all joints}} \text{out}_i$	-2.0
Joint vel limits	$\sum_{\text{all joints}} RELU(\hat{\theta}_i - \hat{\theta}_i^{max})$	-0.1
Torque limits	$\sum_{\text{all joints}} RELU(\hat{\tau}_i - \hat{\tau}_i^{max})$	-0.1
No fly	$\mathbf{1}\{\text{only one feet on ground}\}$	0.25
Feet lateral distance	$\lvert y_{\text{left foot}}^B - y_{\text{right foot}}^B \rvert - d_{min}$	2.5
Feet slip	$\sum_{feet} \lvert \mathbf{v}_i^{\text{toot}} \rvert * \sim \mathbf{1}_{\text{new contact}}$	-0.25
Feet ground parallel	$\sum_{feet} Var(H_i)$	-2.0
Feet contact force	$\sum_{feet} RELU(F_i^z - F_{th})$	-2.5×10^{-4}
Feet parallel	$Var(D)$	-2.5
Contact momentum	$\sum_{feet} \lvert v_i^z * F_i^z \rvert$	-2.5×10^{-4}

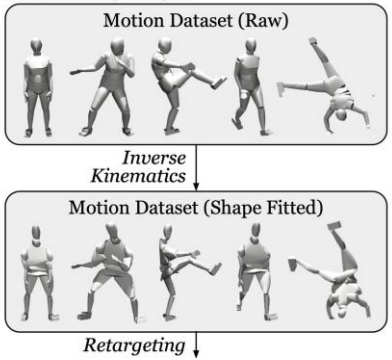
Long et al. 2024

Motivation

Previous works do imitation, but on limited graphics motion datasets with **no terrain or environment**

Why? we need vision!

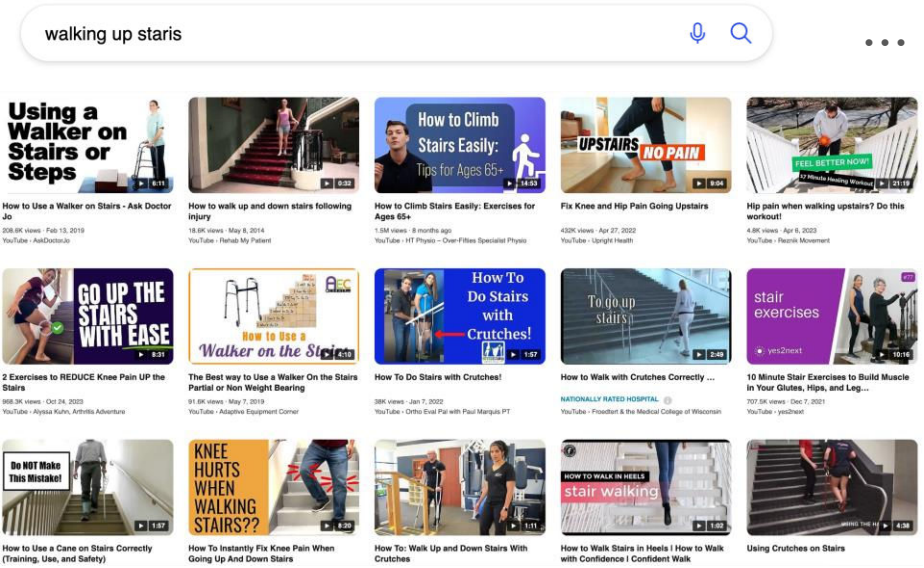
But there's not diverse and realistic data for robot to learn vision-policy!



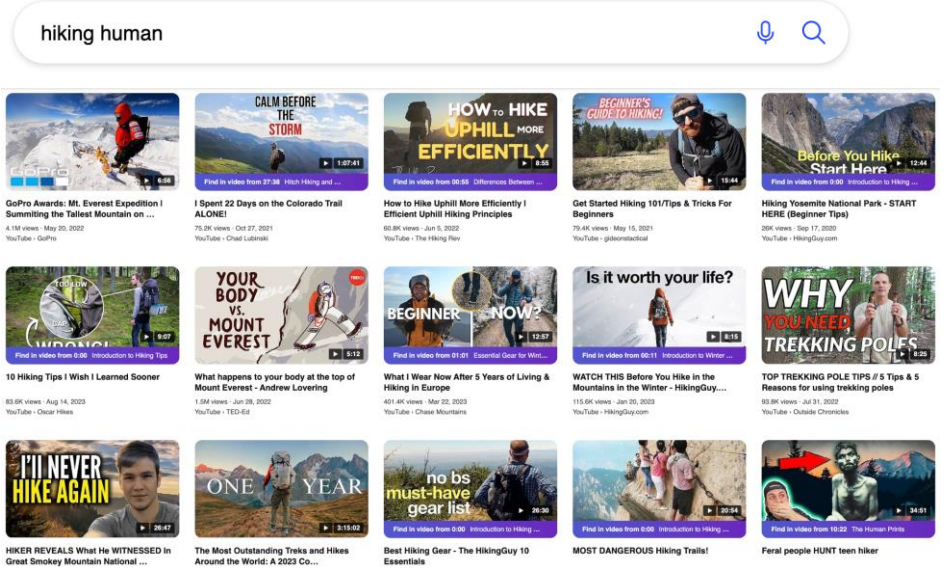
He et al. 2024

Motivation

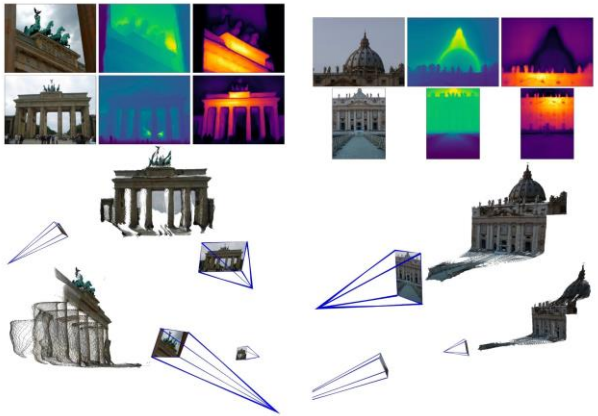
Learn from **Videos from Internet**: diverse human motion, diverse environment



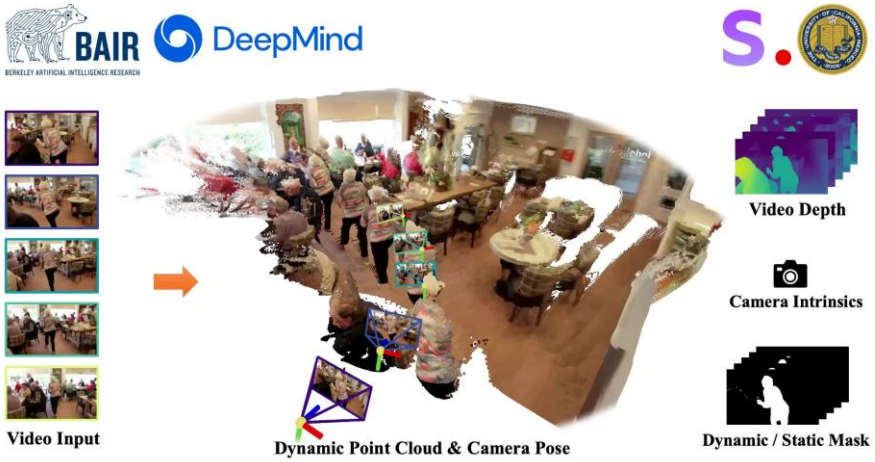
...



Motivation



DUSt3R

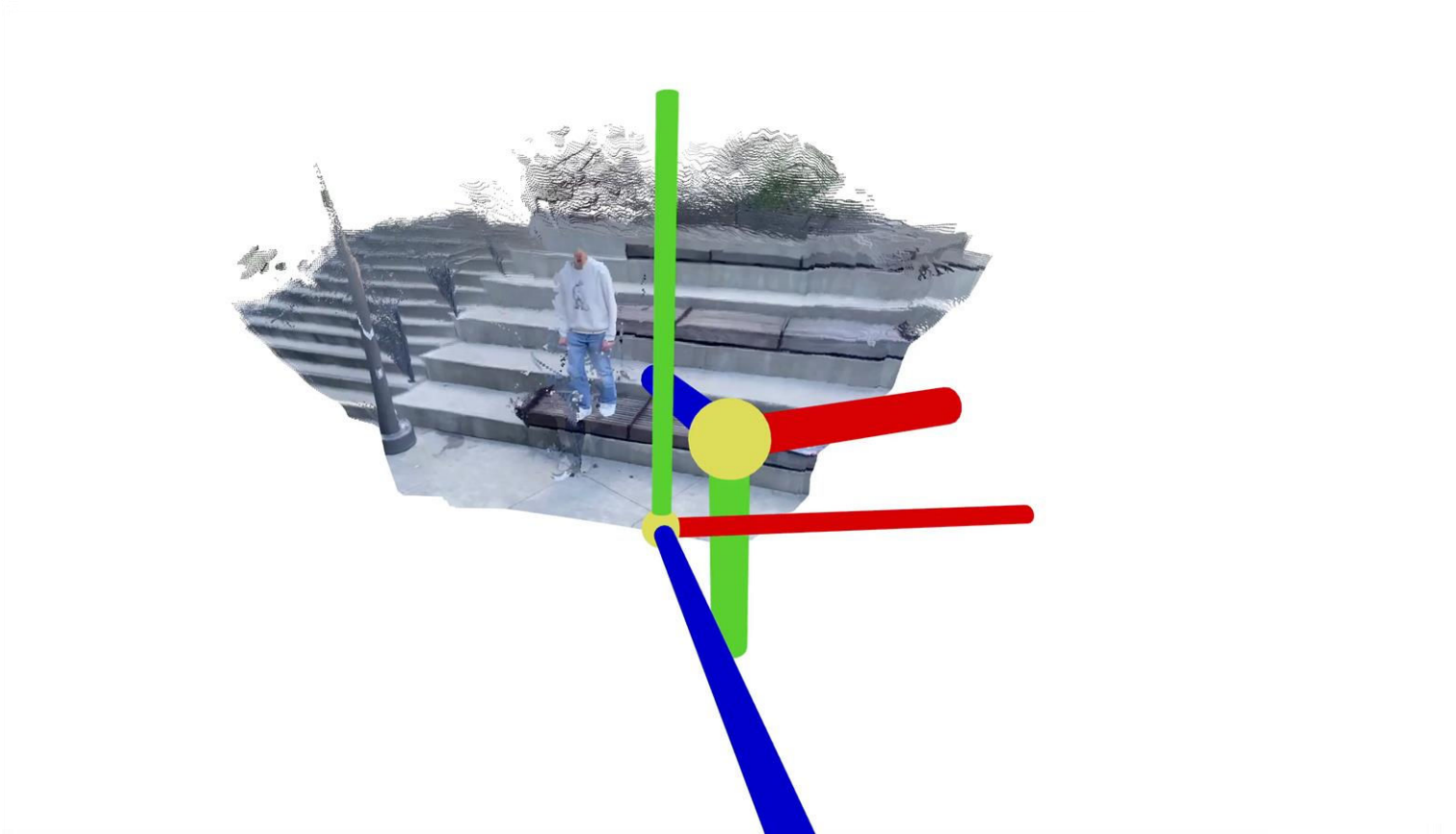


MonST3R

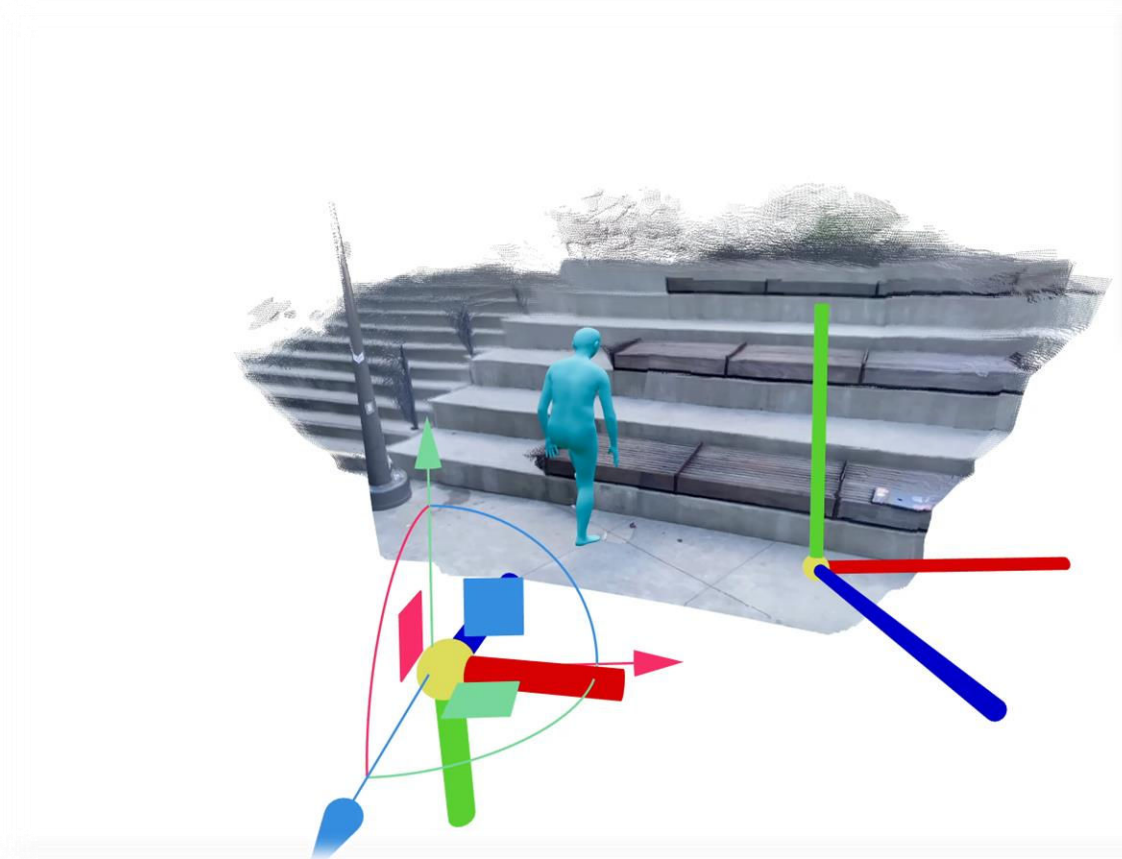
Human SfM

Monocular Video Input

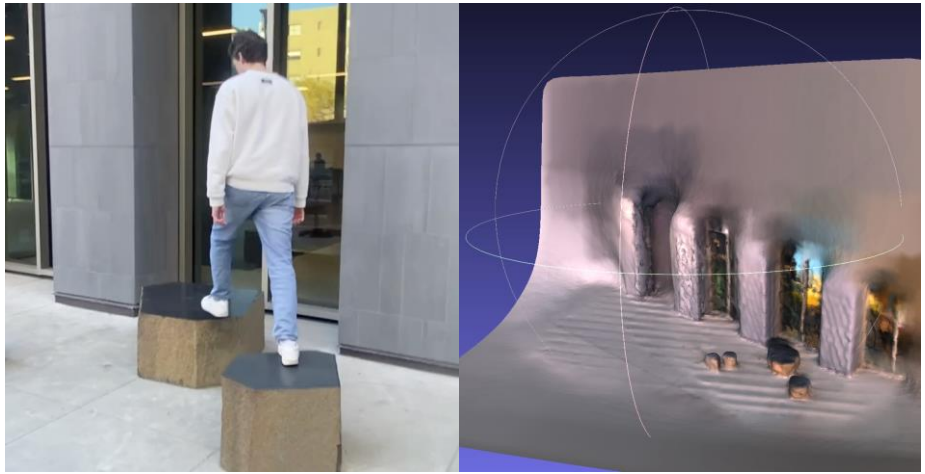
Reconstructed Human and Environment



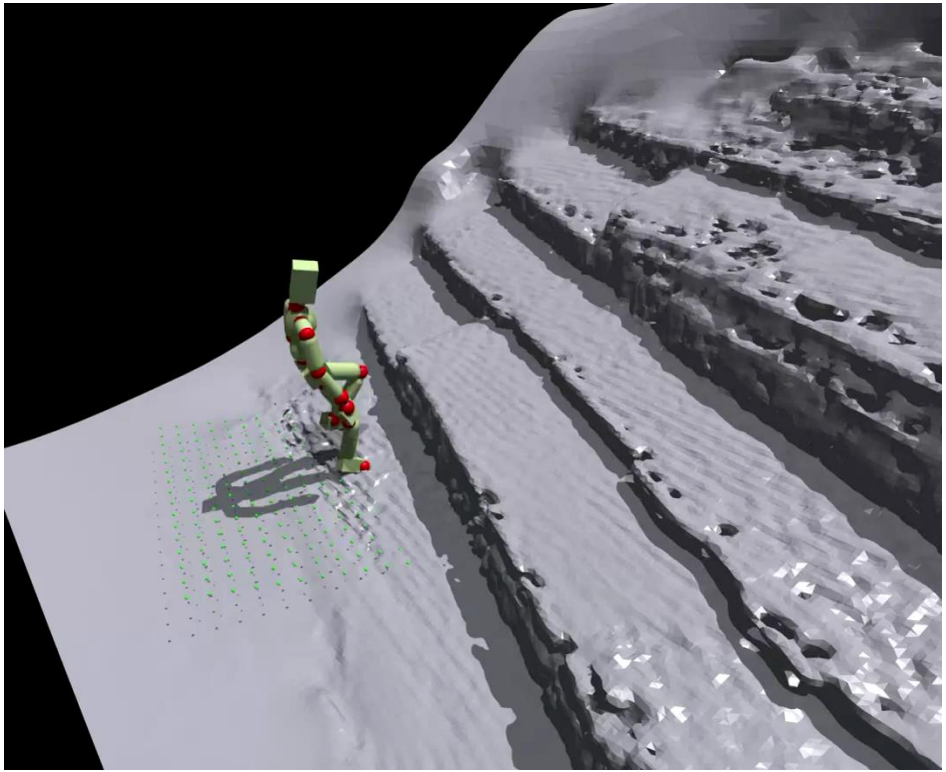
Fit Human Mesh to the PointClouds



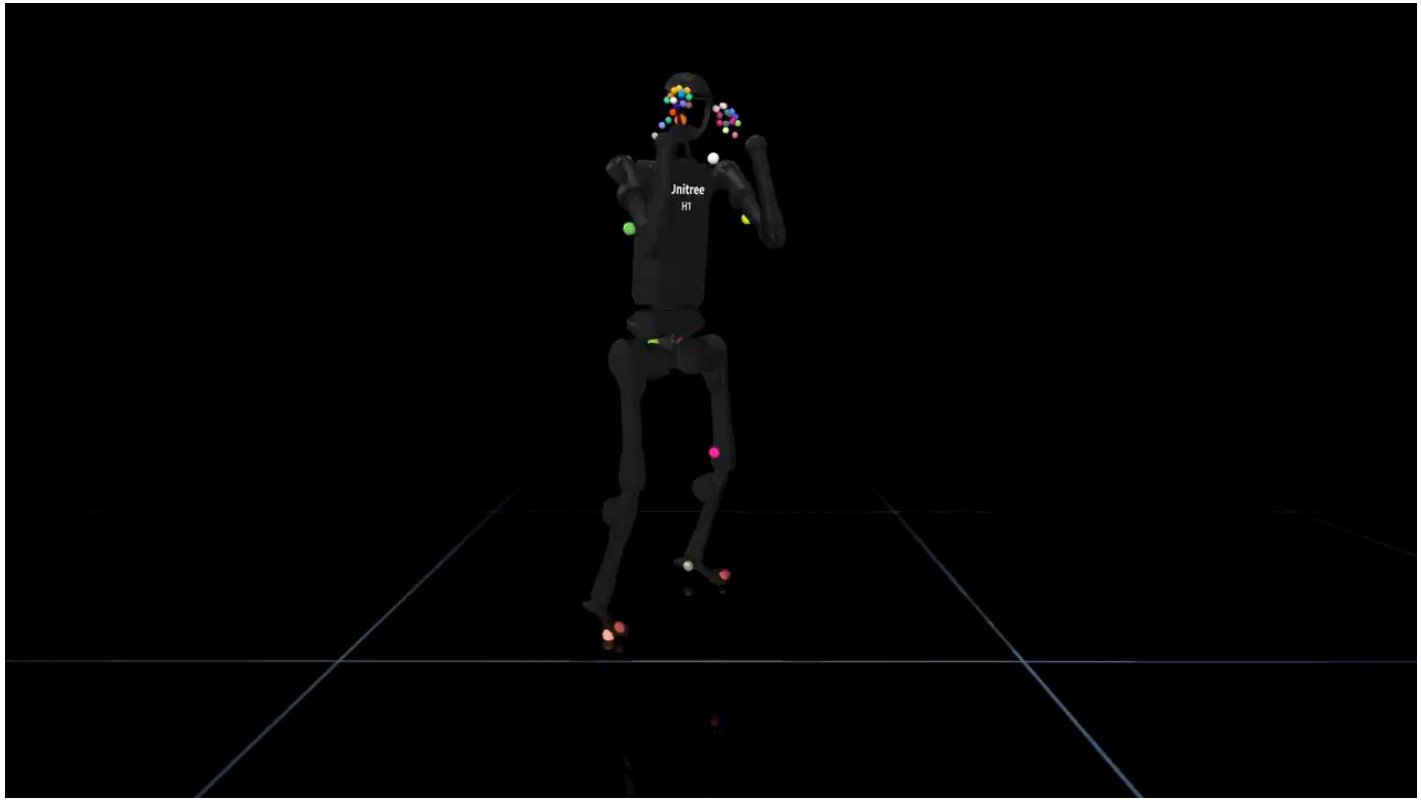
Meshification of the Environment



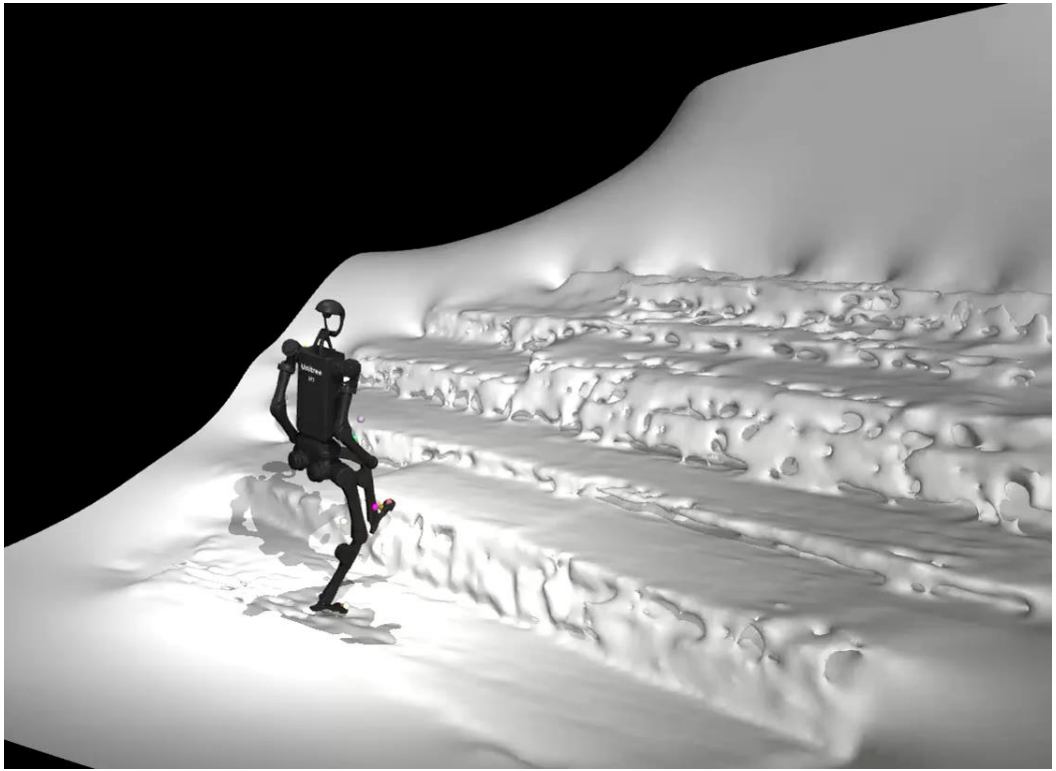
Replaying SMPL Motion with Mesh



Retargeting



Retargeted motion with terrain



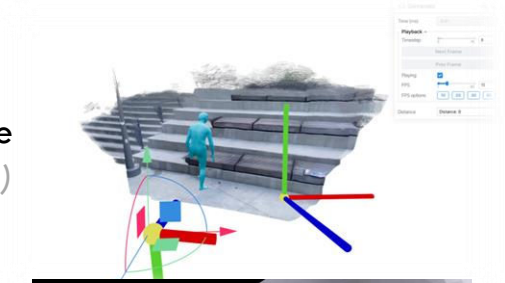
Learning a policy to imitate the retargeted motion

Recap

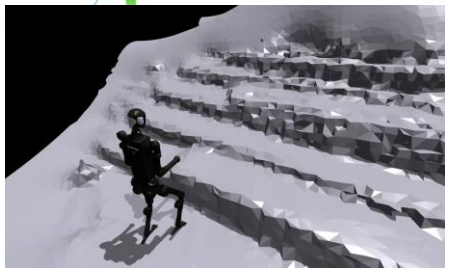
- Current vision pipelines allow for surprisingly good 3D structure from monocular video
- Can leverage this for
  - 3D environment reconstruction
  - Human reconstruction
- Allows for mesh reconstruction and human retargeting
- Can do policy learning on top of the recovered data

Input
(monocular video)

Intermediate
(SMPL + Mesh)



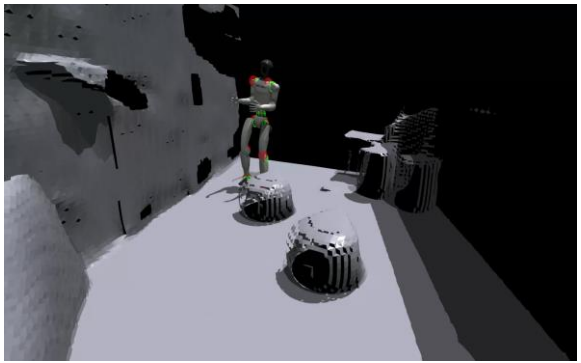
Output
(tracking policy)

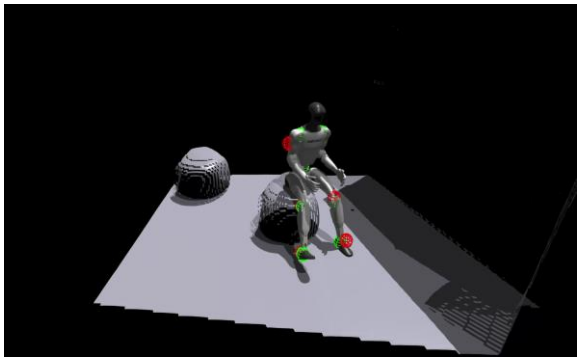


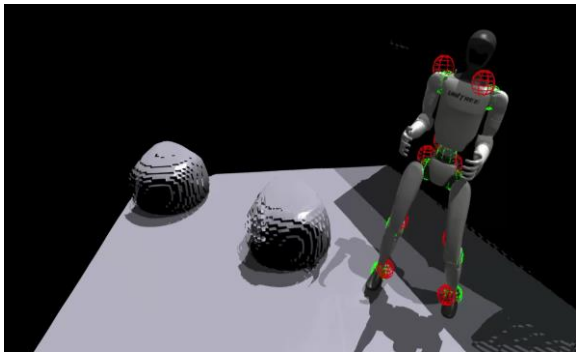
More Results

Input Video

Learn in Sim







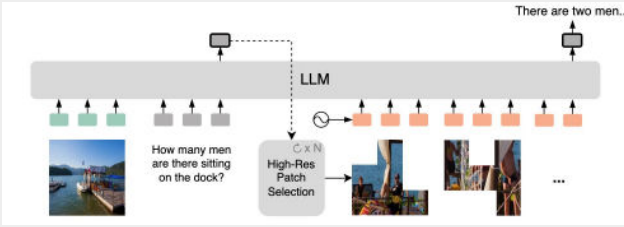
LLMs from Text to Vision and Robotics and back…

- Are LLMs Grounded? … *Surprisingly so!*
- Reducing VLM Hallucination with REVERSE Retrospective Sampling
- Efficient Scaling of VLMs to 4K Resolution with via PS3
- Visual Tokens for Non-linguistic Generation (ViLex)
- Navigation World Models with CDiT
- **4D Reconstruction for Humanoid Robotics with StarTr4K, ARM4R, and VideoMimic**

(Efficient) LLMs from Text to Vision and Robotics and back…

Prof. Trevor Darrell

UC Berkeley



$$J(s_{\tau:\tau+m}, s^*, a_{\tau:\tau+m-1}) = \text{Distance to Goal} + \text{Action Constraints} + \text{State Constraints}$$

States Goal Actions