A Tale of Two Realities

Bridging Physical Worlds with Interactable Digital Twins for Embodied Robots

Baoxiong Jia

BIGAI

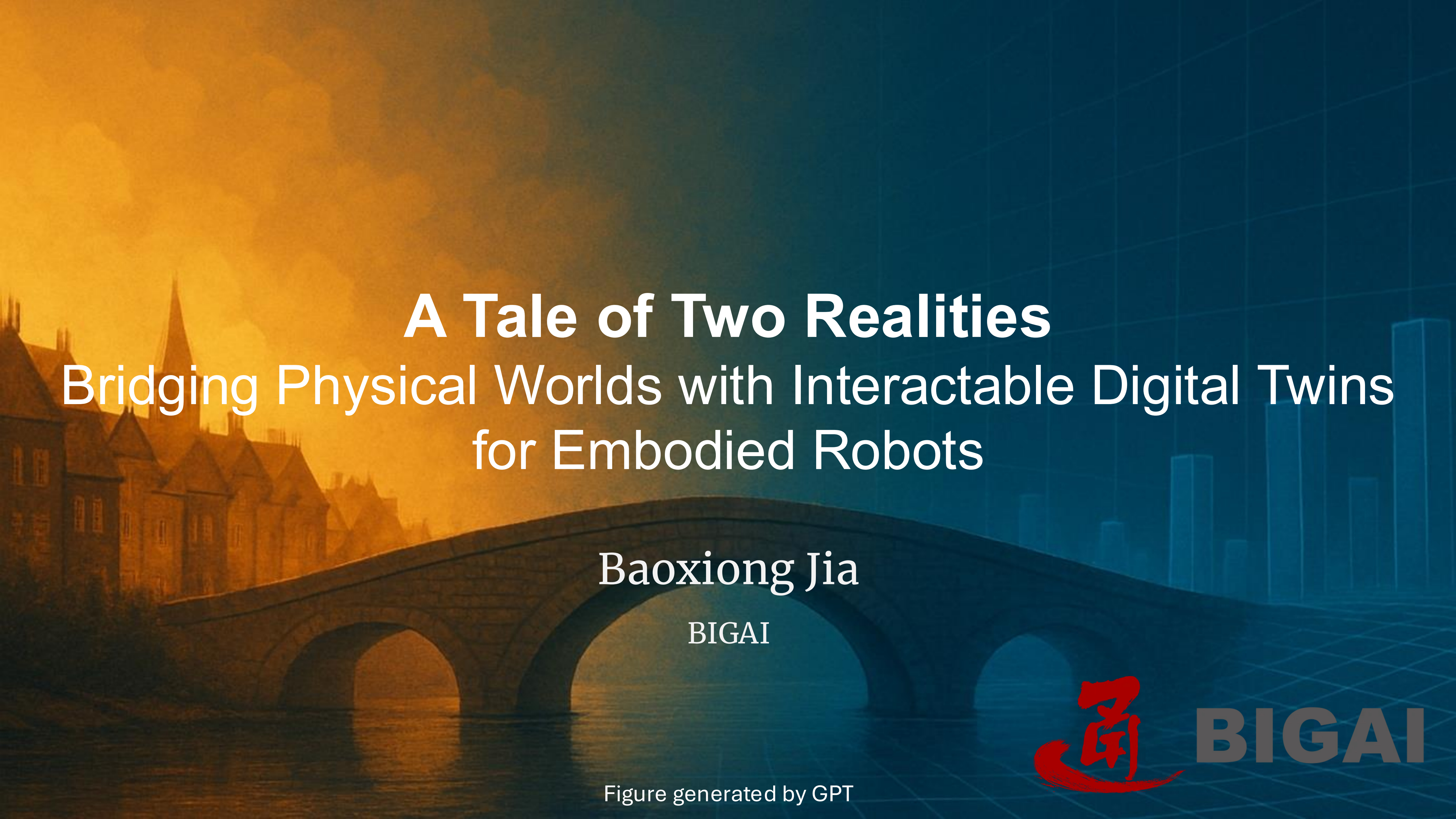


Figure generated by GPT

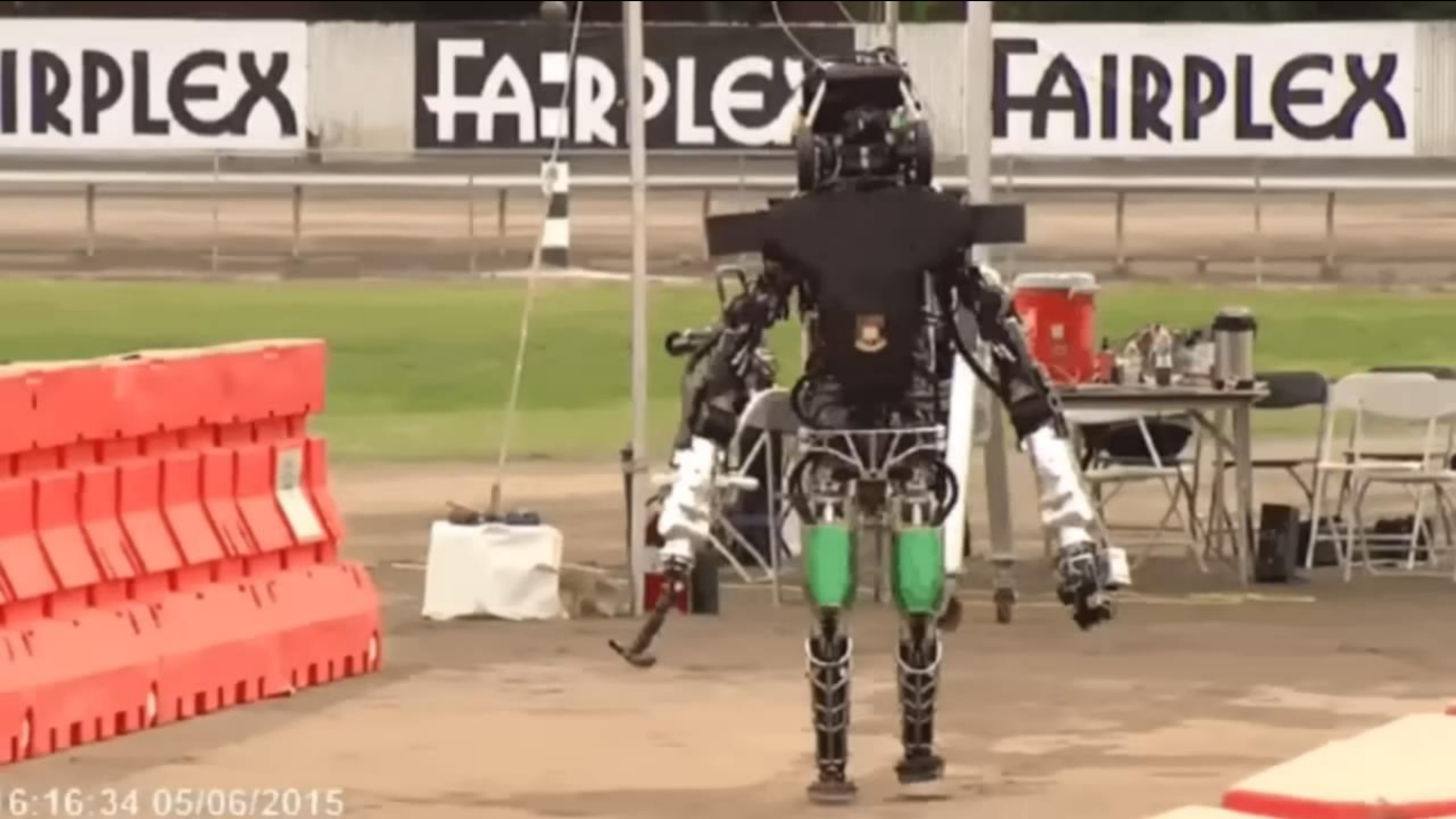
About me

buzz-beater.github.io

Peking University
B.S. in CS
2014-2018

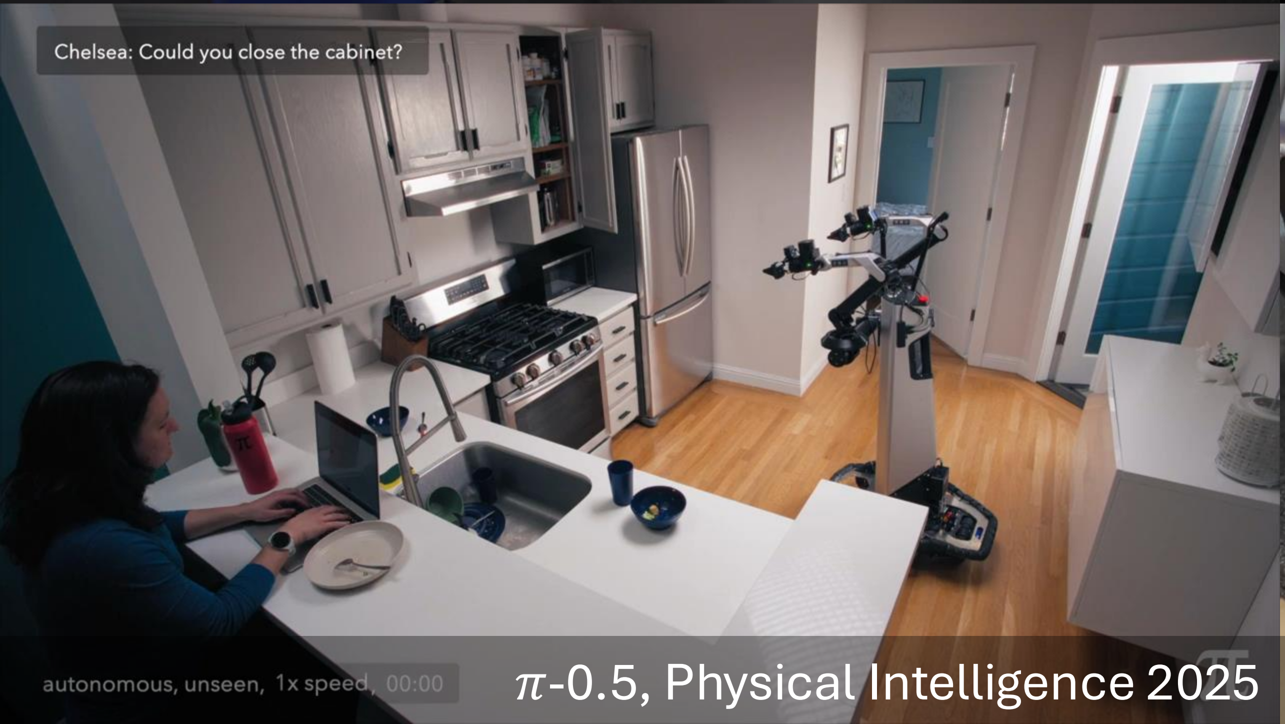
UCLA
Ph.D. in CS
2018-2022

BIGAI
Research Scientist
2022-Present


FAIRPLEX
FAIRPLEX
FAIRPLEX
6:16:34 05/06/2015

Loco-Manipulation, Boston Dynamics 2025

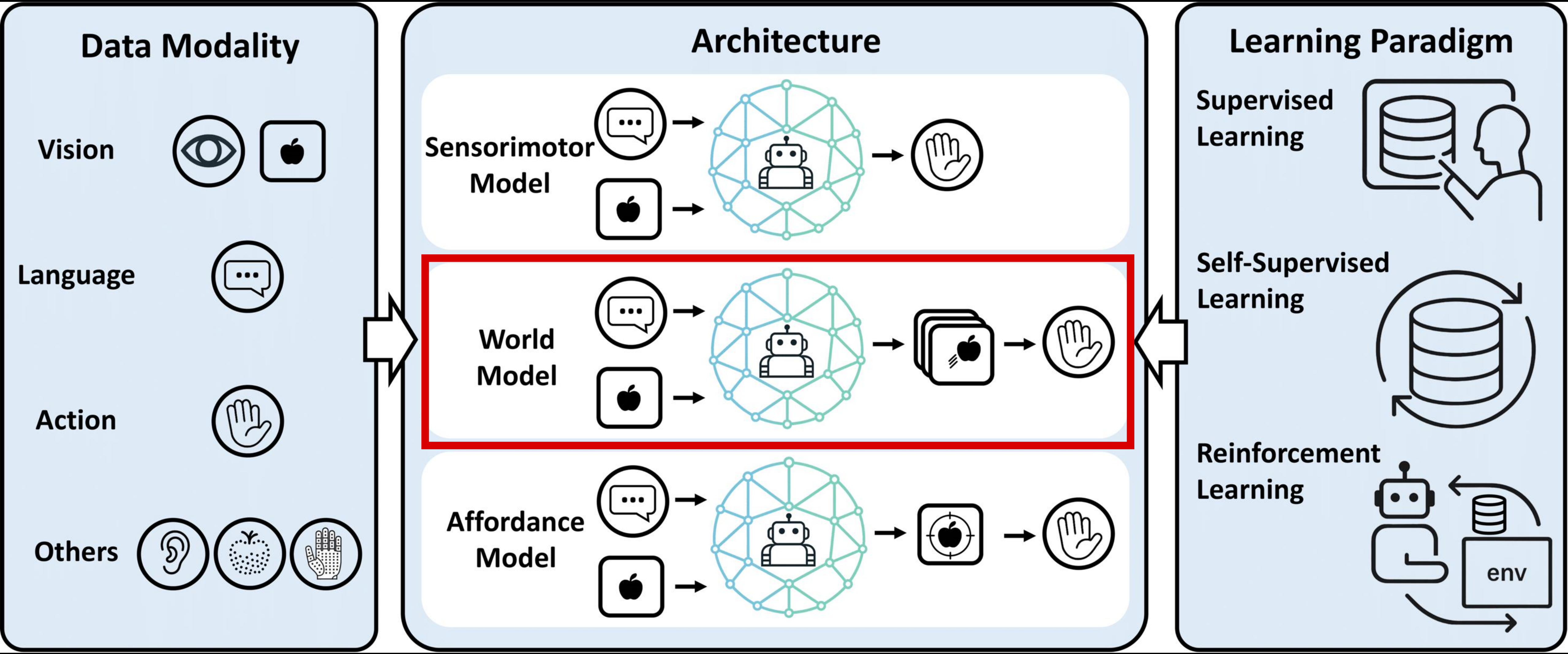
Introducing Helix, Figure 2025



π-0.5, Physical Intelligence 2025


Early Preview of Model Capabilities, Generalist 2025

VLA for Embodied Robots



Kawaharazuka et al., Vision-Language-Action Models for Robotics: A Review Towards Real-World Applications, IEEE Access 2025

World Models

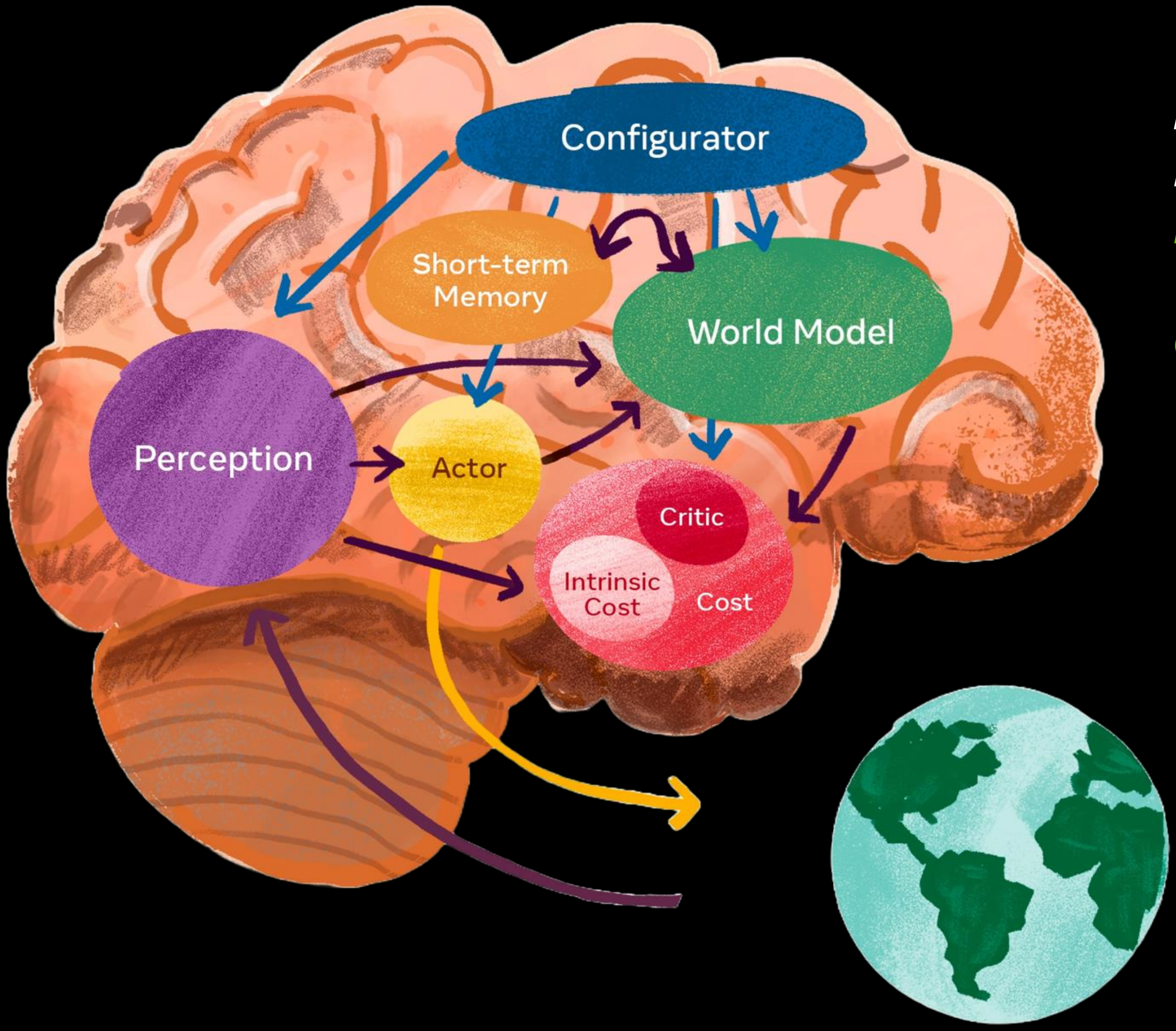


Figure credit: JEPA Blog, Meta 2022

"If the organism carries a small-scale model of external reality and of its own possible actions within its head, it is able to try out various alternatives, conclude which is the best of them, react to future situations before they arise, utilise the knowledge of past events in dealing with the present and future."

— Kenneth Craik (1943)

- **Integration of perception and action**
  - **The model must encode states and possible actions**
- **Prediction, reasoning and planning**
  - **The model functions as an internal simulator for anticipating outcomes and guiding decisions**
- **Efficient representation and generalization**
  - **Retains essential structures to predict the future and generalize past experience**

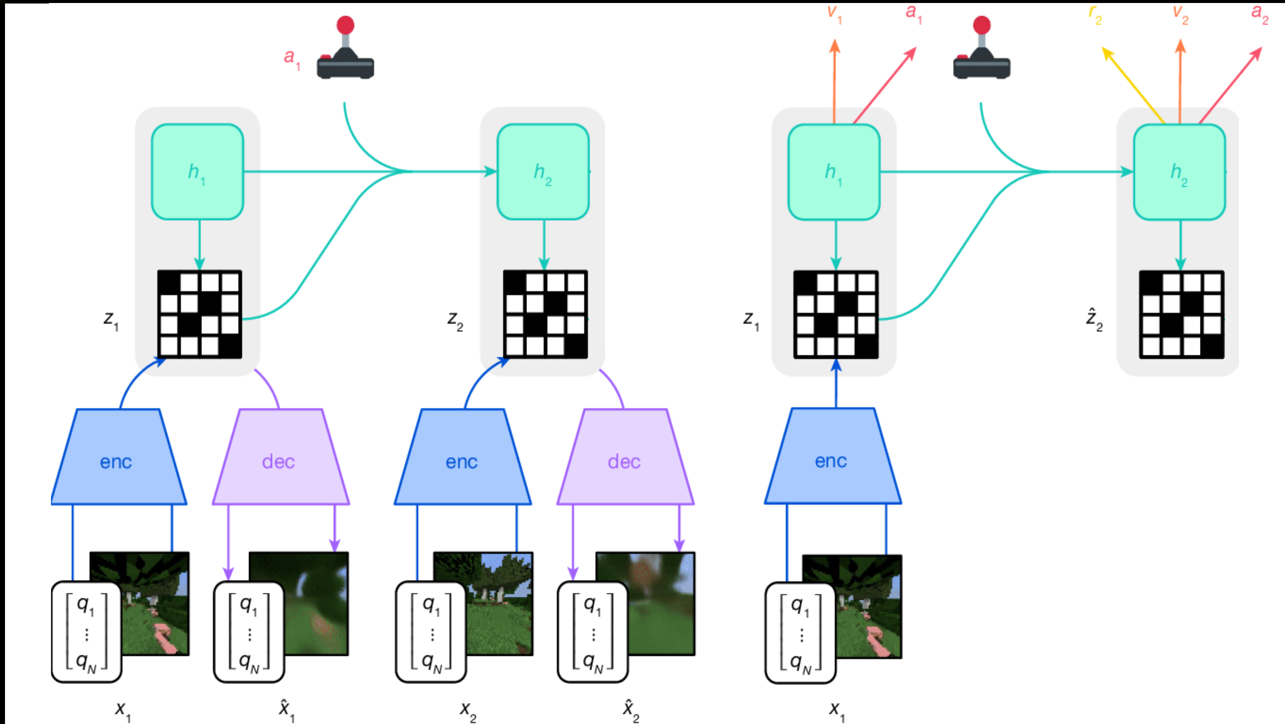
Model-based RL

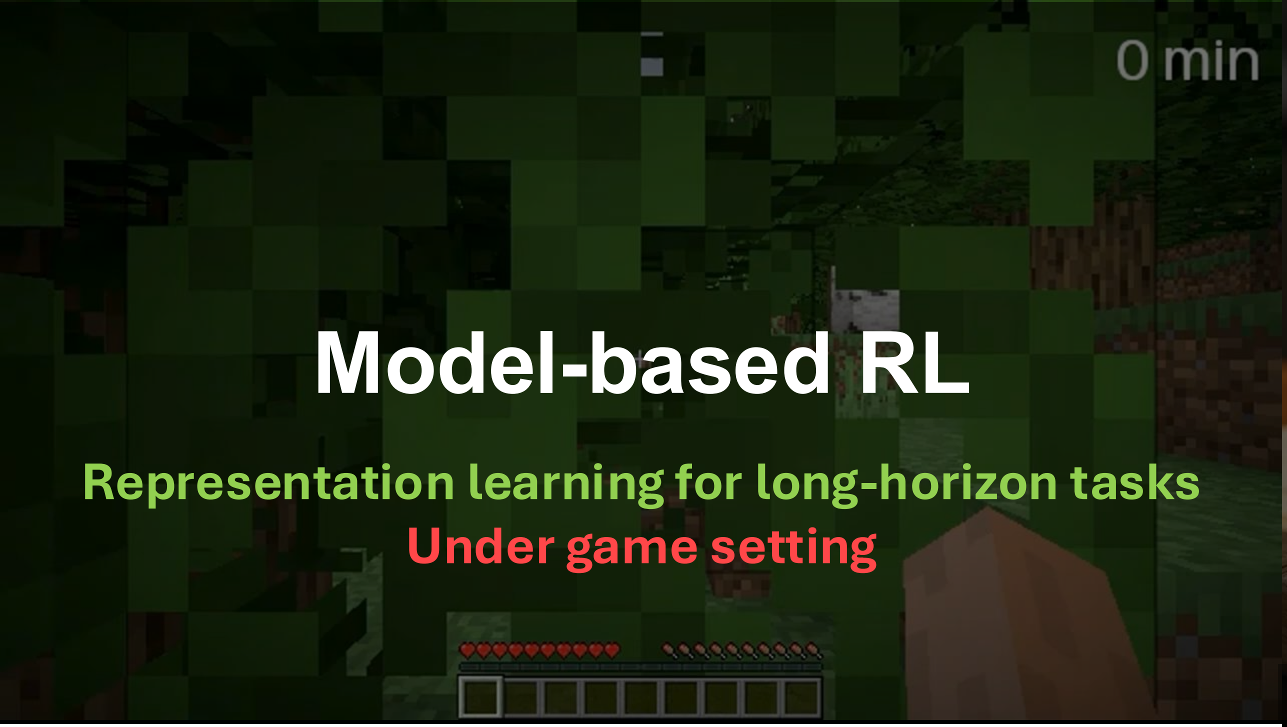
Representation learning for long-horizon tasks

Under game setting

0 min

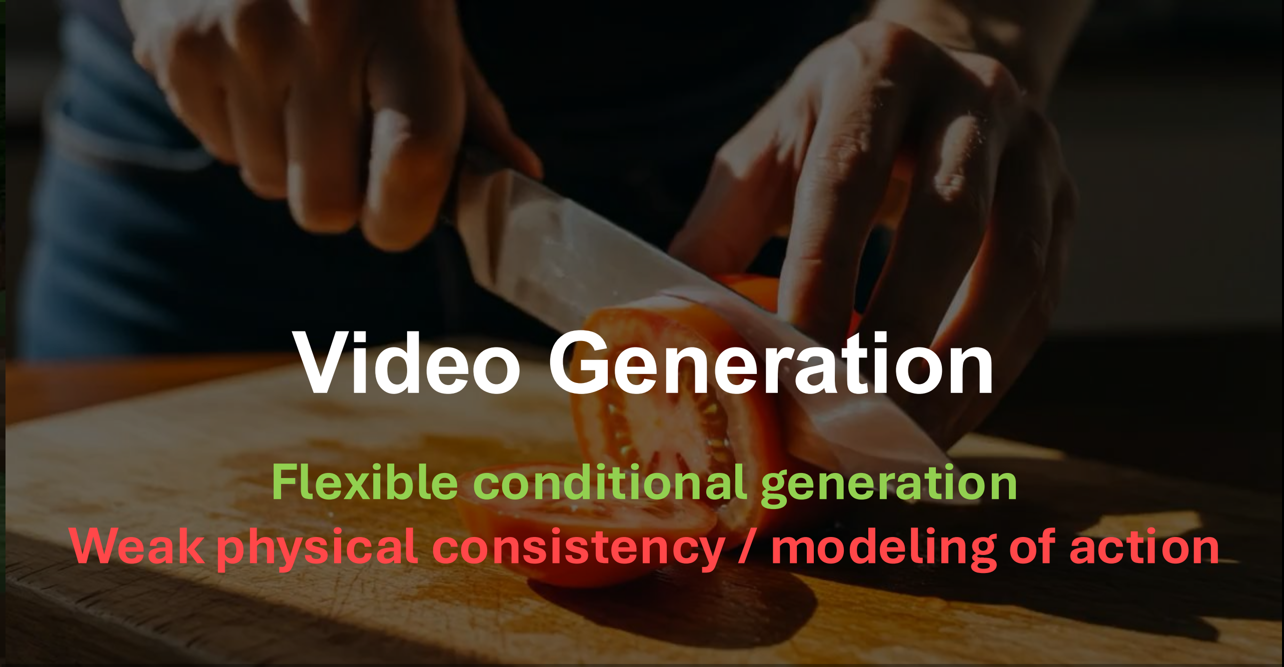
Dreamer 4, Google DeepMind 2025





Model-based RL

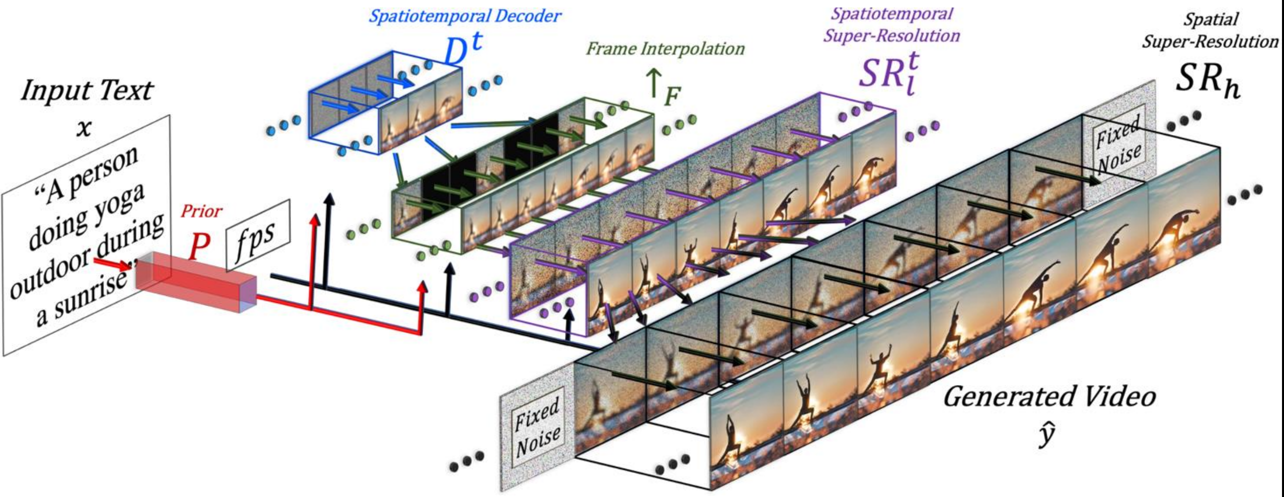
Representation learning for long-horizon tasks
Under game setting



Video Generation

Flexible conditional generation
Weak physical consistency / modeling of action

Veo 3.1, Google Deepmind 2025



Model-based RL

0 min

Representation learning for long-horizon tasks

Under game setting

Video Generation

Flexible conditional generation

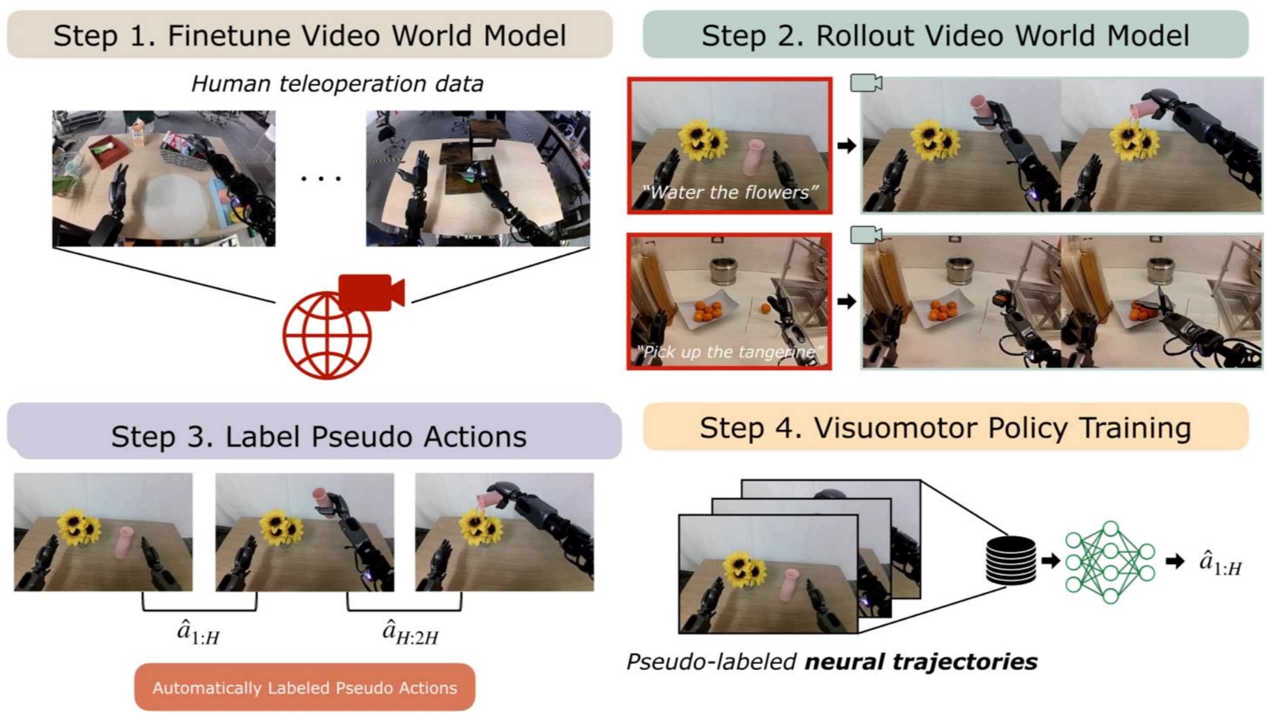
Weak physical consistency / modeling of action

Dream to Generate

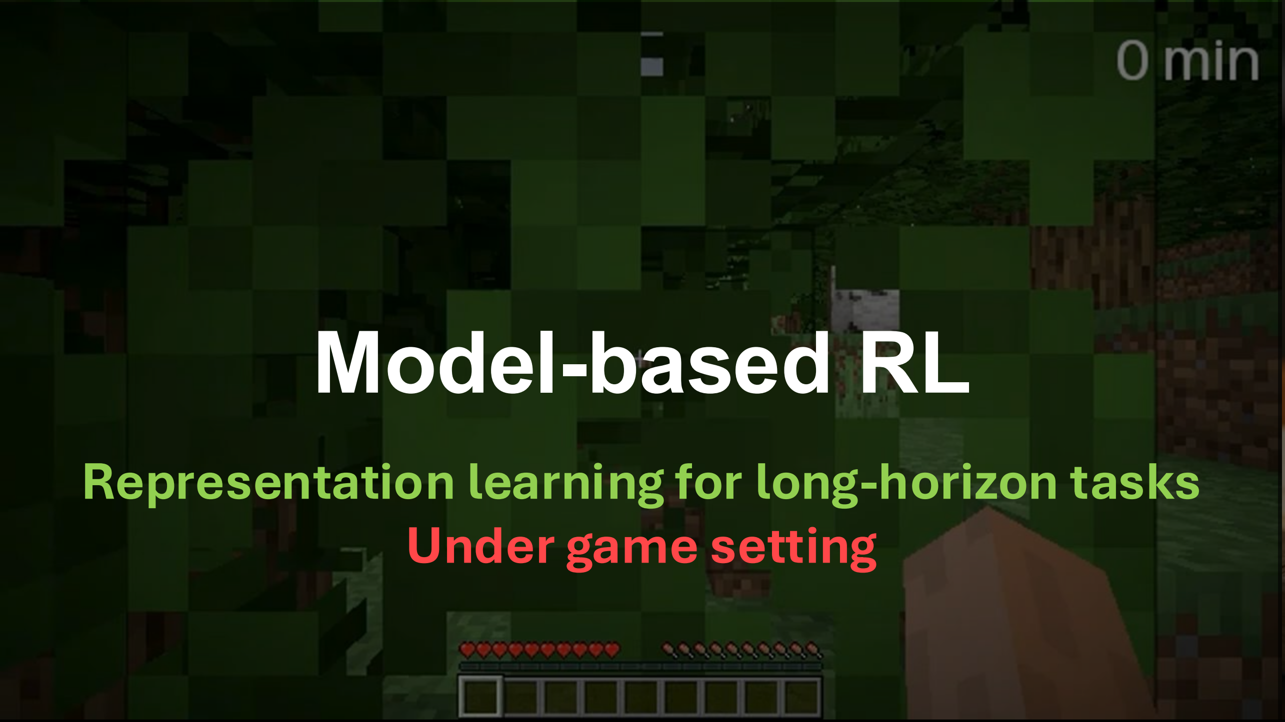
Latent Action Learning

Aligning video generation with latent actions

Limited by the view point



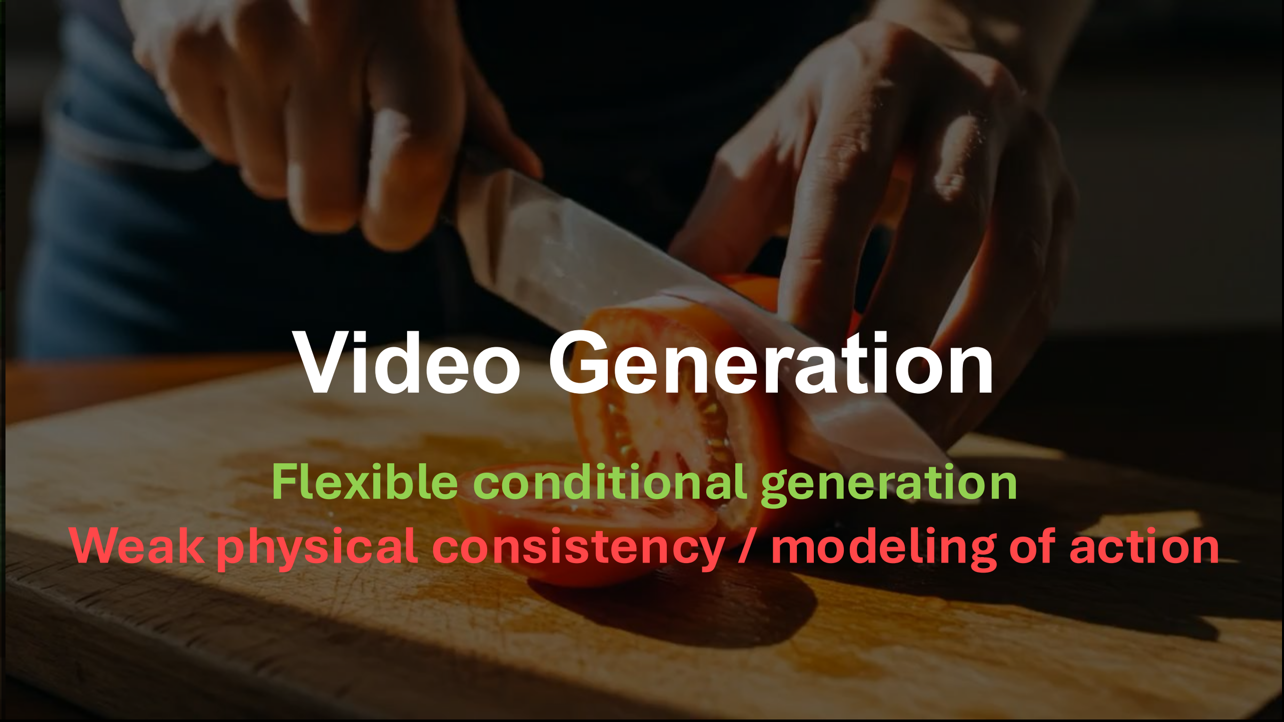
DreamGen, NVIDIA GEAR 2025



Model-based RL

Representation learning for long-horizon tasks

Under game setting



Video Generation

Flexible conditional generation

Weak physical consistency / modeling of action

Dream to Generate

Latent Action Learning

Aligning video generation with latent actions

Limited by the view-point

Spatial Representations

World modeling with 3D Gaussians

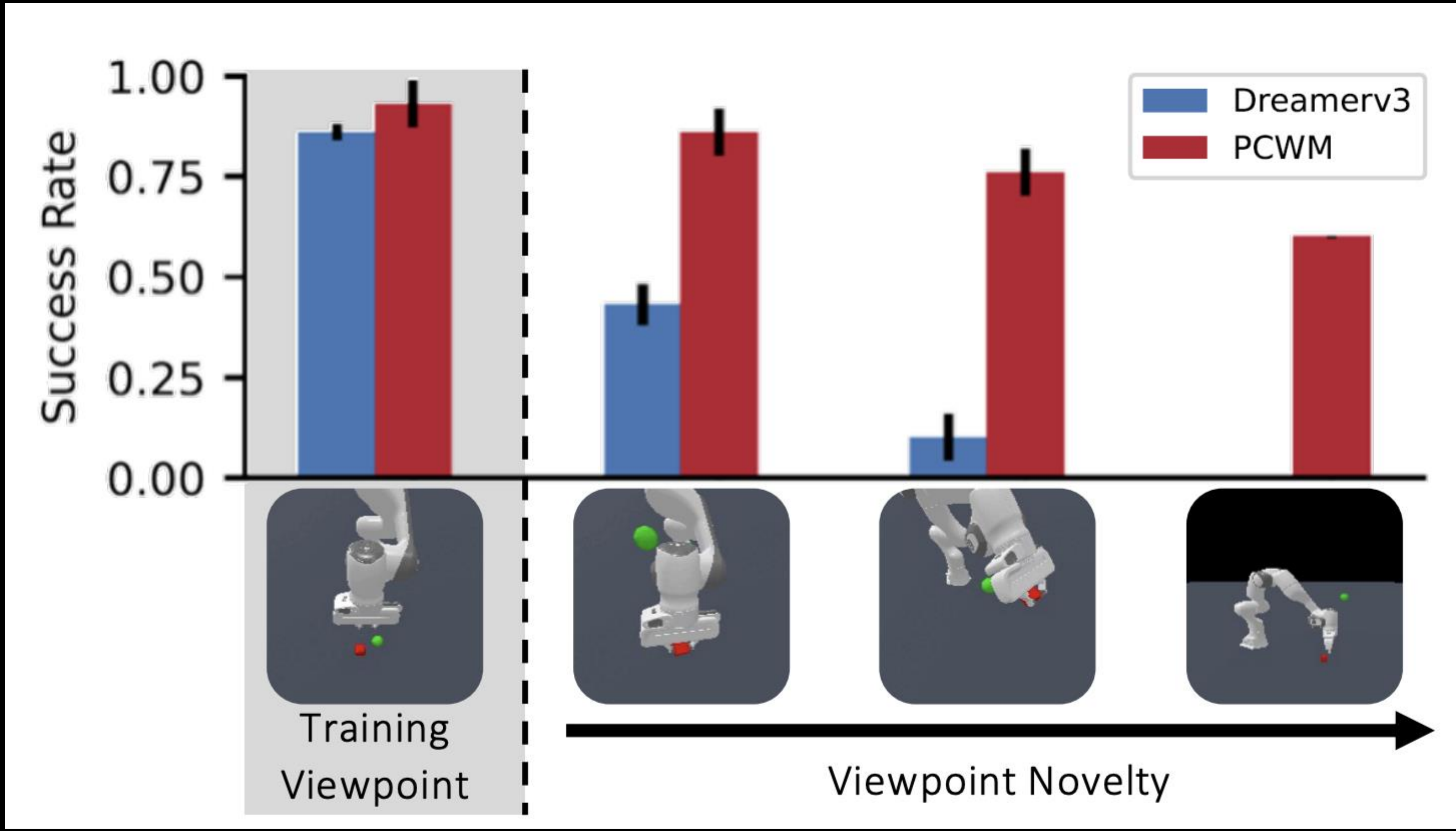
Interactiveness for robot manipulation?

Marble, WorldLabs 2025

3D World Models?

Because of depth estimation challenge, tele-op must follow protocols

3D World Models?



3D helps policy learning, but requires additional sensors (RGB-D)

Peri et al., Point Cloud Models Improve Visual Robustness in Robotic Learners, ICRA 2024

A Naïve Idea: Use Feed-Forward 3D Gaussians as a Flexible and Efficient Representation







Guanxing Lu* | **Baoxiong Jia*** ✉️ | Puhao Li* | Yixin Chen | Ziwei Wang | Yansong Tang ✉️ | Siyuan Huang ✉️

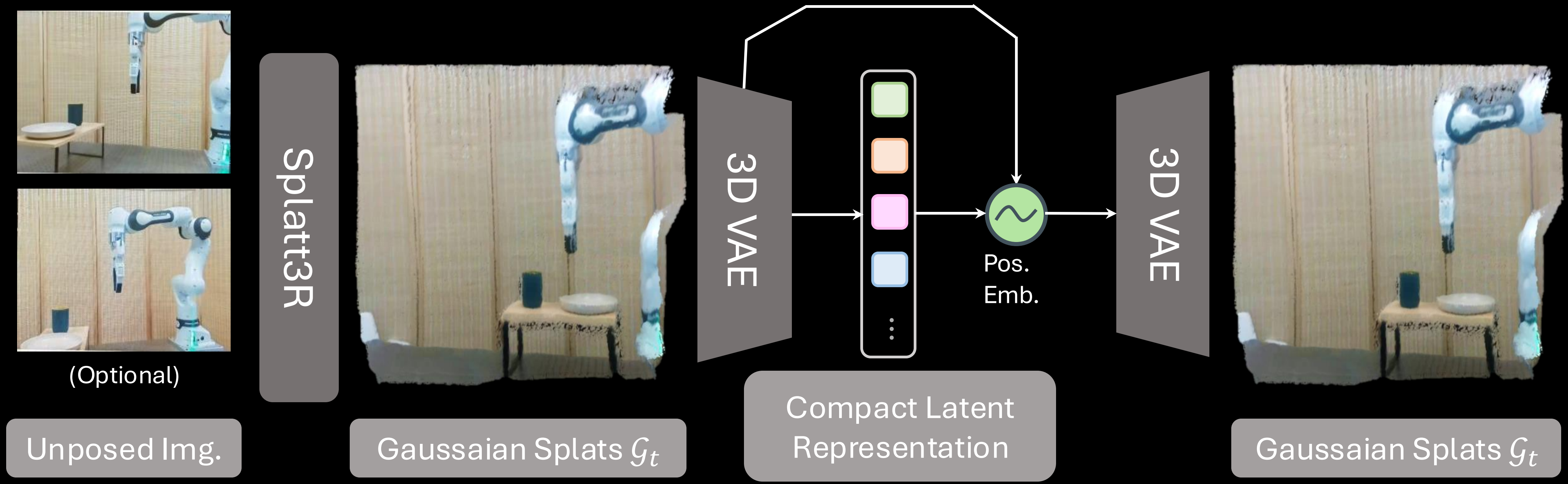
GWM: Towards Scalable Gaussian World Model for Robotic Manipulation

https://gaussian-world-model.github.io

ICCV OCT 19-23, 2025 HONOLULU HAWAII

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

Encoding 3D Gaussians into Latent Space

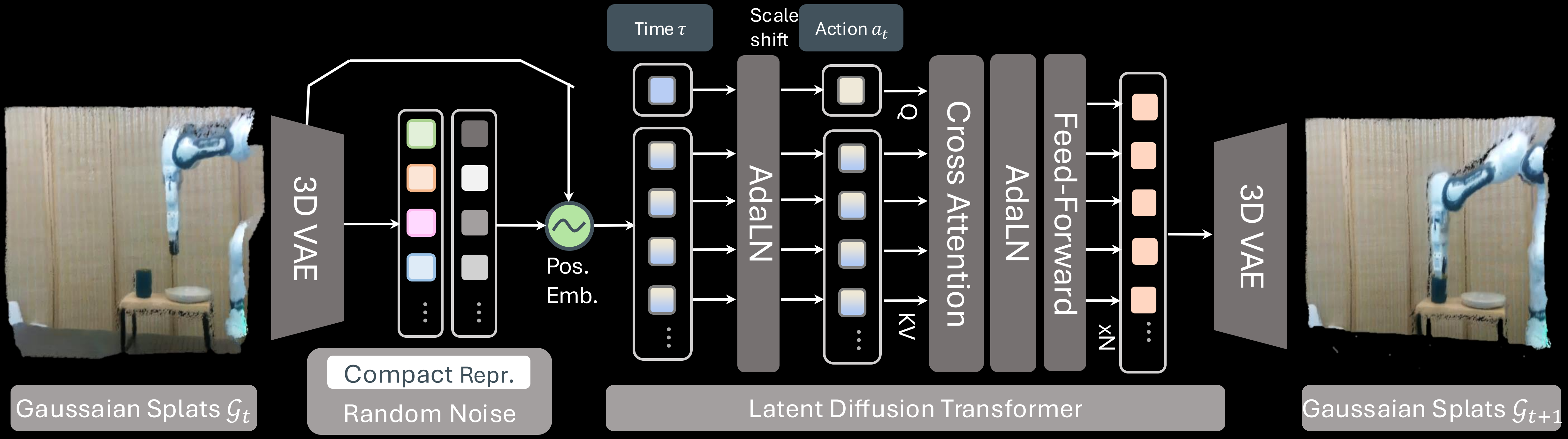


Feed-Forward 3D Gaussian Reconstruction

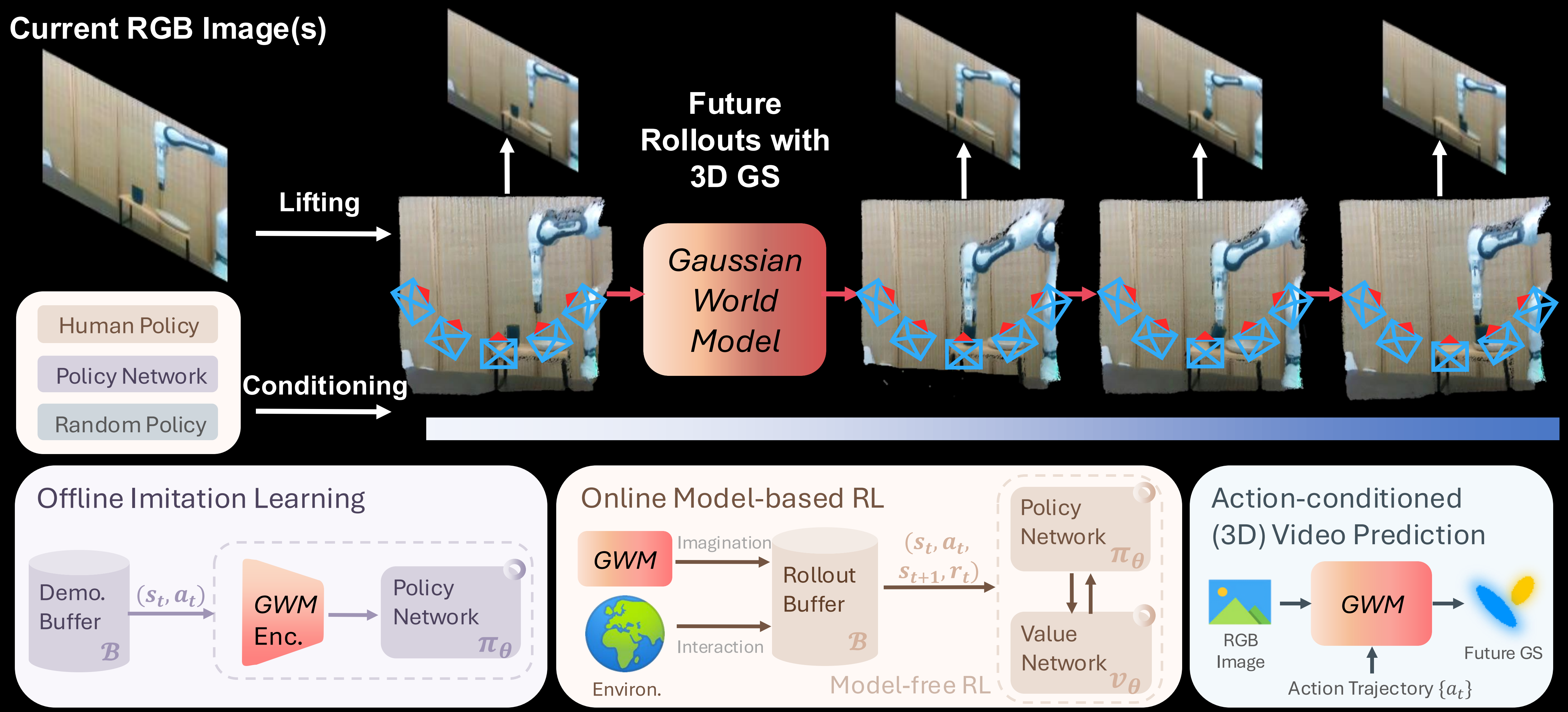
FPS-based Subsampling Query-based Encoding

Rendering / Geometry Supervision

GWM: Gaussian World Model



DiT-based Dynamics Learning and Prediction



Current RGB Image(s)
Lifting
Future Rollouts with 3D GS
Gaussian World Model
Human Policy
Policy Network
Random Policy
Conditioning
Offline Imitation Learning
Demo. Buffer
$\mathcal{B}$
(s_t, a_t)
GWM Enc.
Policy Network
π_θ
Online Model-based RL
GWM
Imagination
Rollout Buffer
$\mathcal{B}$
(s_t, a_t, s_{t+1}, r_t)
Policy Network
π_θ
Value Network
v_θ
Interaction
Environ.
Model-free RL
Action-conditioned (3D) Video Prediction
RGB Image
GWM
Future GS
Action Trajectory $\{a_t\}$

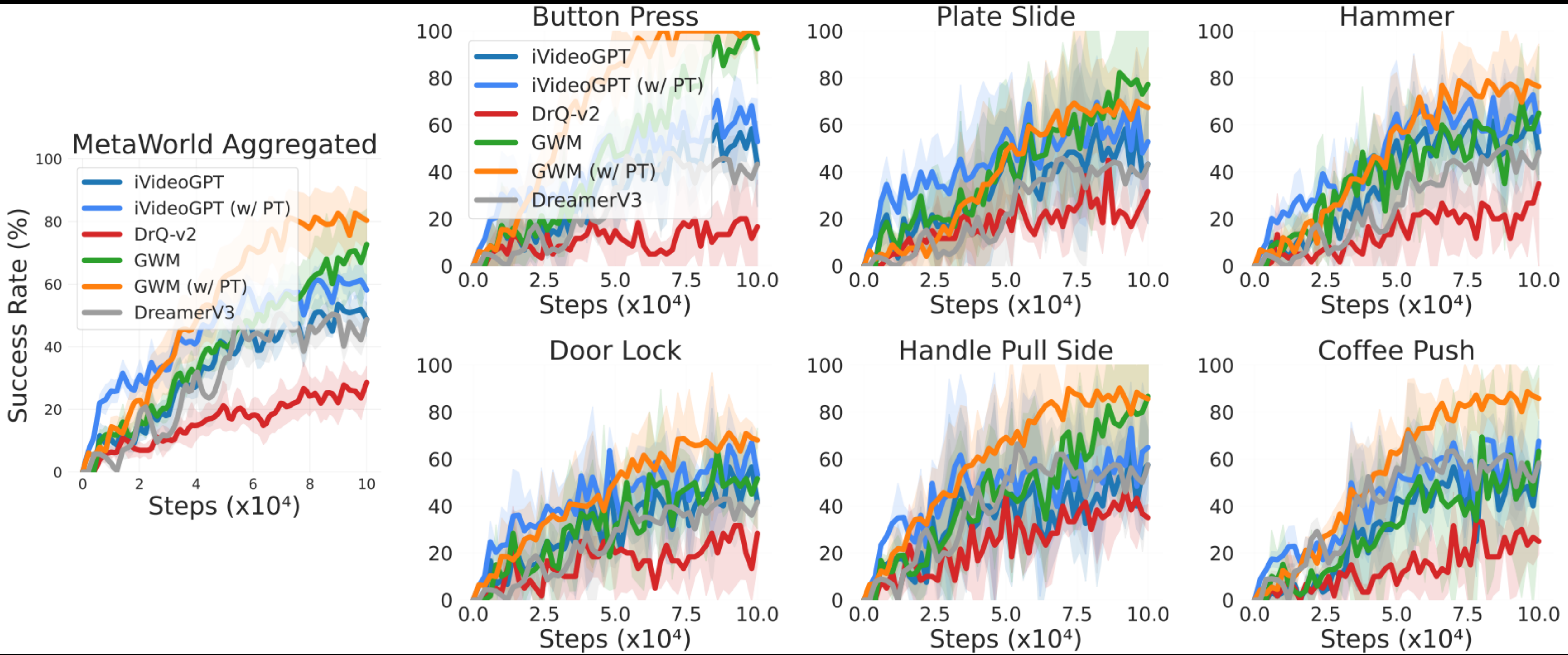
Method	PnP CabToCounter H-50	PnP CabToCounter G-3000	PnP CounterToCab H-50	PnP CounterToCab G-3000	PnP CounterToMicrowave H-50	PnP CounterToMicrowave G-3000	PnP CounterToSink H-50	PnP CounterToSink G-3000	PnP CounterToStove H-50	PnP CounterToStove G-3000	PnP MicrowaveToCounter H-50	PnP MicrowaveToCounter G-3000
BC-transformer	2	18	6	28	2	18	2	44	2	6	2	8
GWM	18	32	4	22	14	44	20	38	2	18	20	26
Δ	+16	+14	-2	-6	+12	+26	+18	-6	0	+12	+18	+18

Method	PnP SinkToCounter H-50	PnP SinkToCounter G-3000	PnP StoveToCounter H-50	PnP StoveToCounter G-3000	Open SingleDoor H-50	Open SingleDoor G-3000	Open DoubleDoor H-50	Open DoubleDoor G-3000	Close DoubleDoor H-50	Close DoubleDoor G-3000	Close SingleDoor H-50	Close SingleDoor G-3000
BC-transformer	8	42	6	28	46	50	28	48	28	46	56	94
GWM	22	38	18	44	58	62	28	42	50	58	54	90
Δ	+14	-4	+12	+16	+12	+12	0	-6	+22	+12	-2	-4

Method	Open Drawer H-50	Open Drawer G-3000	Close Drawer H-50	Close Drawer G-3000	TurnOn Stove H-50	TurnOn Stove G-3000	TurnOff Stove H-50	TurnOff Stove G-3000	TurnOn SinkFaucet H-50	TurnOn SinkFaucet G-3000	TurnOff SinkFaucet H-50	TurnOff SinkFaucet G-3000
BC-transformer	42	74	80	96	32	46	4	24	38	34	50	72
GWM	56	90	80	90	46	80	22	40	52	48	44	66
Δ	+14	+16	0	-6	+14	+24	+18	+16	+14	+14	-6	-6

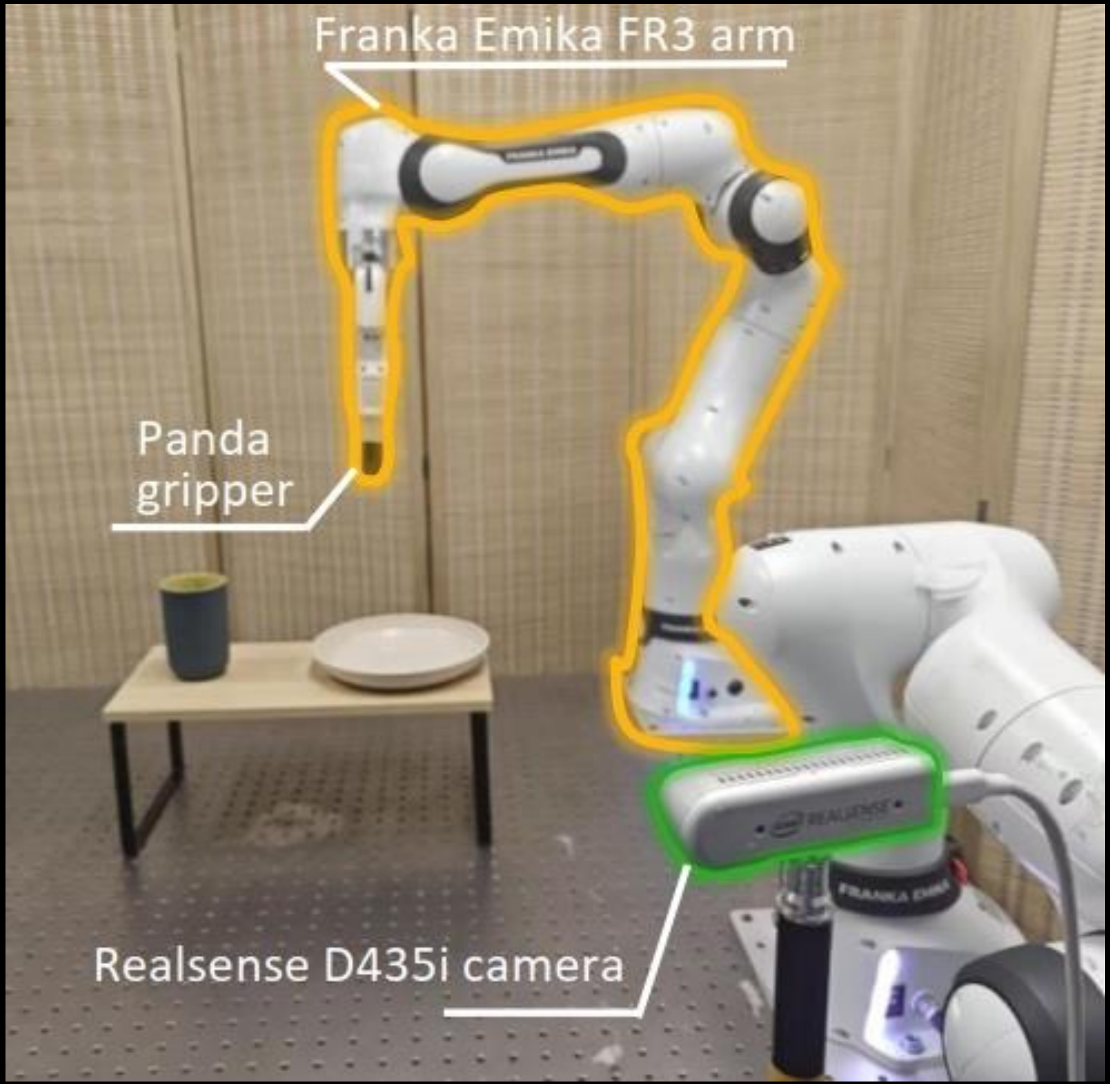
Method	Turn SinkSpout H-50	Turn SinkSpout G-3000	CoffeePress Button H-50	CoffeePress Button G-3000	TurnOn Microwave H-50	TurnOn Microwave G-3000	TurnOff Microwave H-50	TurnOff Microwave G-3000	CoffeeServe Mug H-50	CoffeeServe Mug G-3000	CoffeeSetup Mug H-50	CoffeeSetup Mug G-3000
BC-transformer	54	96	48	74	62	90	70	60	22	34	0	12
GWM	72	90	76	90	64	84	70	54	36	50	16	28
Δ	+18	-6	+28	+16	+2	-6	0	-6	+14	+16	+16	+16

GWM for Online Model-based RL



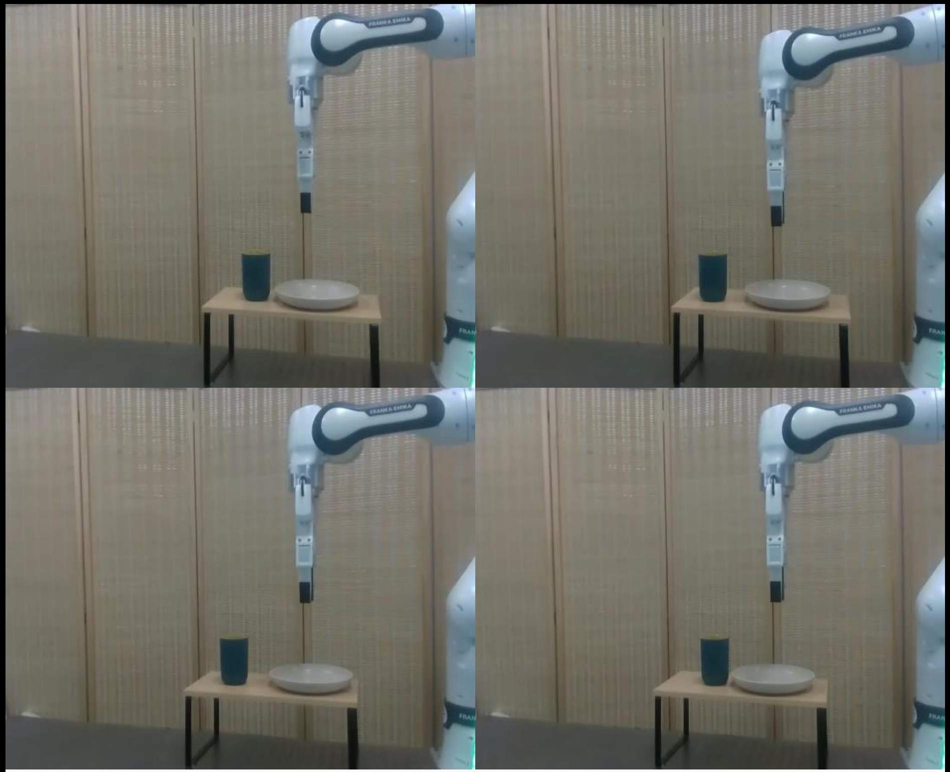
Additional reward learning on top of GWM for online RL

GWM for Real-World Robot Manipulation



Diffusion Policy

DP w/ GWM



Comparsion

FRANKA-PNP	Diffusion Policy	GWM (Ours)
Cup distractor	$6/10$	$\mathbf{7/10}$
Plate distractor	$1/5$	$\mathbf{3/5}$
Table distractor	$0/5$	$\mathbf{3/5}$
Total	$7/20$	$\mathbf{13/20}$

Takeaways

- **Encoding explicit spatial information into world modeling**
  - **Unite world modeling with 3D generation, video generation, multi-view reconstruction, etc.**

Takeaways

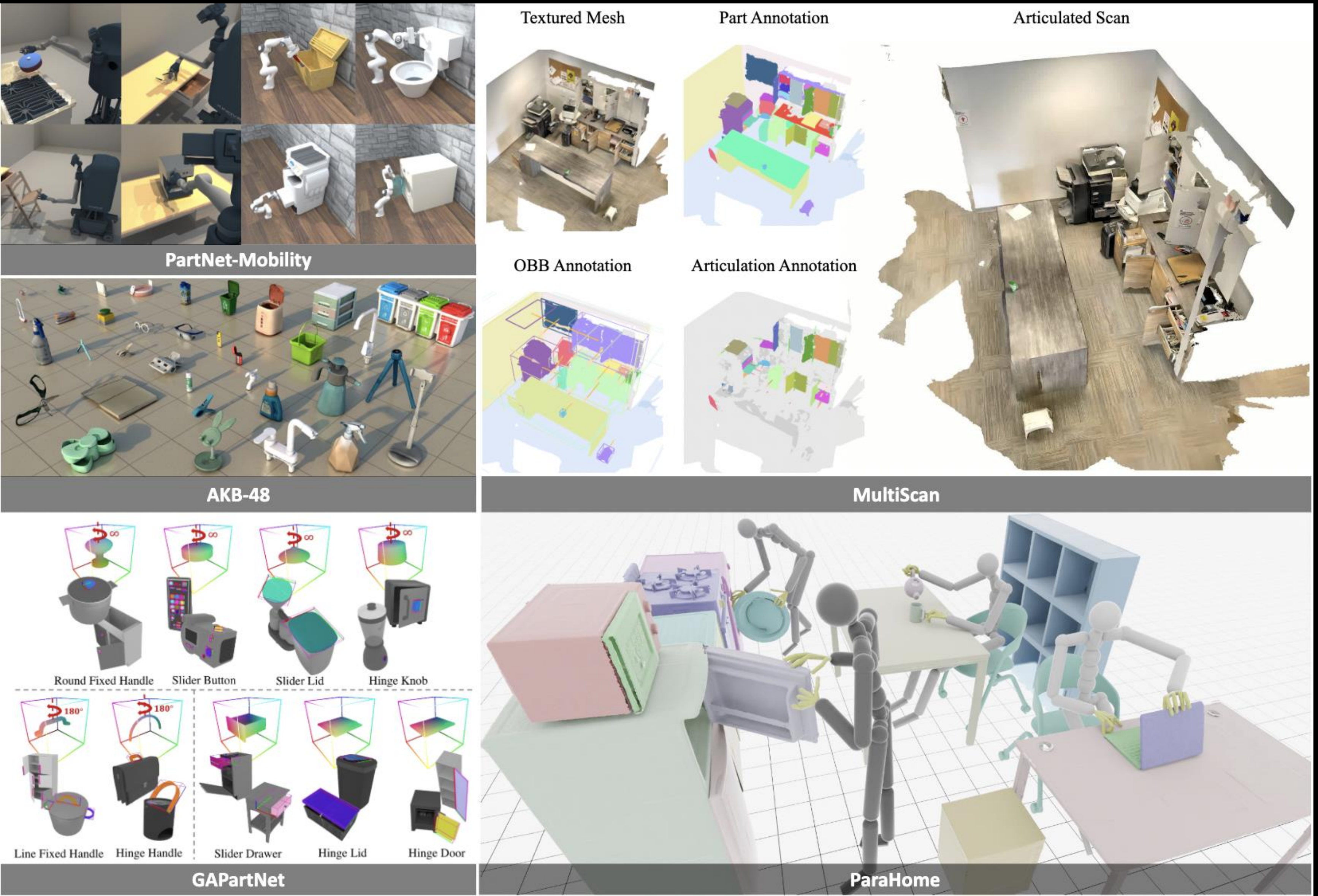
- Encoding explicit spatial information into world modeling
  - Unite world modeling with 3D generation, video generation, multi-view reconstruction, etc.
- Better VLA modeling with world modeling
  - Using the latent representation alone does not fully utilize the predictive power of world models

Takeaways

- ➢ Encoding explicit spatial information into world modeling
  - ❖ Unite world modeling with 3D generation, video generation, multi-view reconstruction, etc.
- ➢ Better VLA modeling with world modeling
  - ❖ Using the latent representation alone does not fully utilize the predictive power of world models
- ➢ Scalable 4D world modeling
  - ❖ Scalability vs. precision still stands as an issue, feed-forward 3D Gaussians still need improvement

Especially for Dynamic Objects

Manipulation Policies involve Dynamic Objects



In reality, we deal with dynamic, articulated objects whose geometry and shape change during interaction, making them difficult to reconstruct

Efficient and Scalable Reconstruction of Articulated Objects from Monocular Video








Yu Liu Baoxiong Jia✉️ Ruijie Lu Chuyue Gan Huayu Chen Junfeng Ni Song-Chun Zhu Siyuan Huang✉️

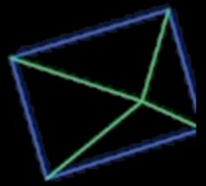
VideoArtGS: Building Digital Twins of Articulated Objects from Monocular Video

https://articulate-gs.github.io

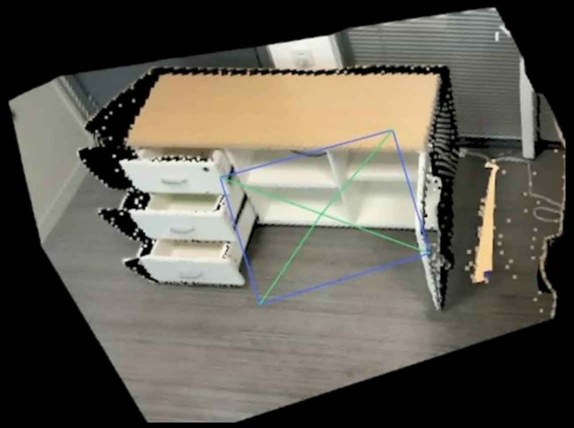
Image Supervision is Ambiguous for Articulation Learning

Key Challenge: The observed pixel motion results from four entangled factors:

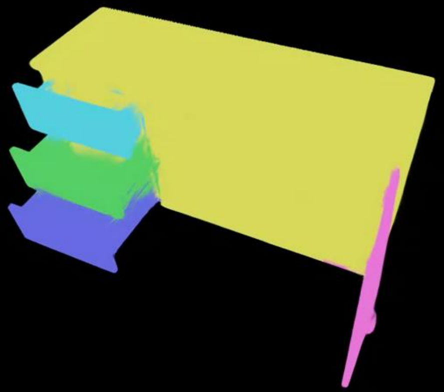
Camera trajectory



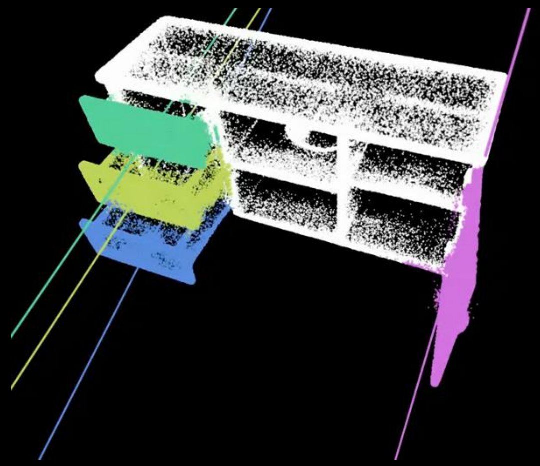
Object Geometry



Part Segmentation

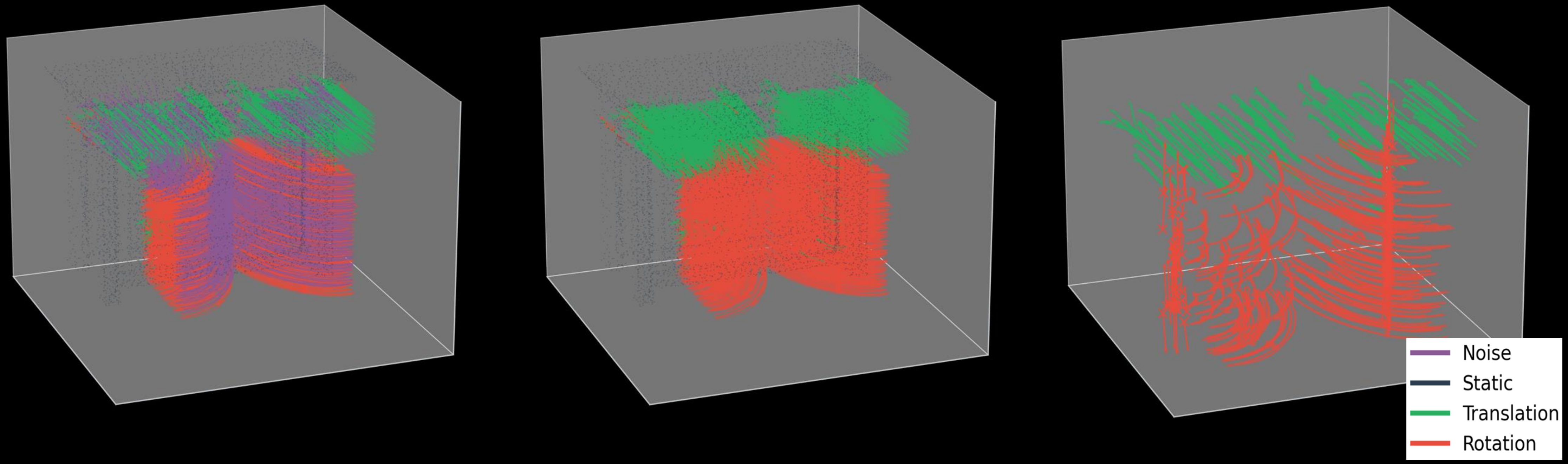


Articulation Dynamics



Provide motion prior from pre-trained tracking models

Key Insights: Analyze noisy 3D tracks to provide robust initialization and optimization signals



Filtering noise and estimate articulation parameters

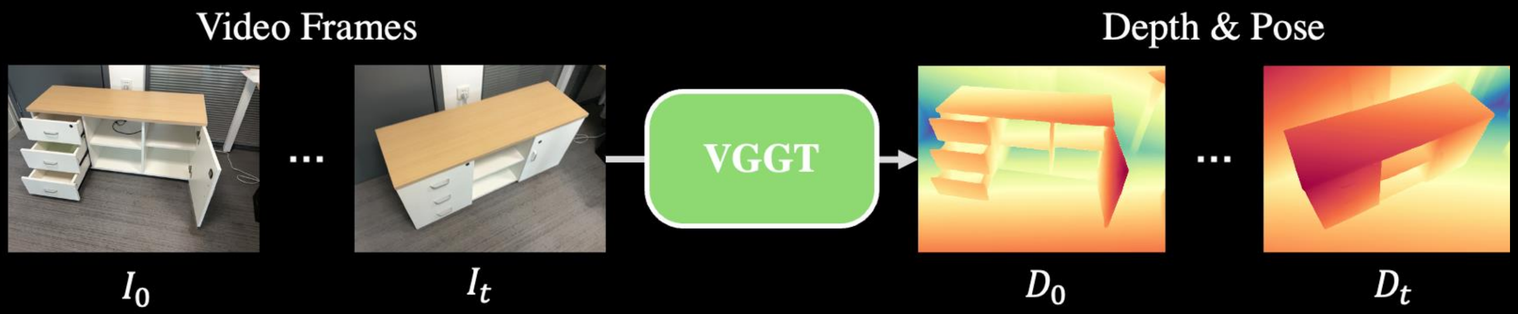
Camera, Depth, Tracks Estimation

Video Frames

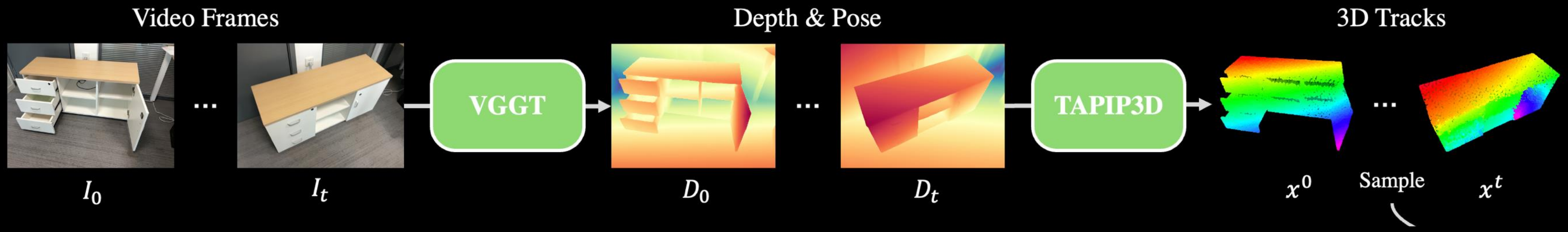
...

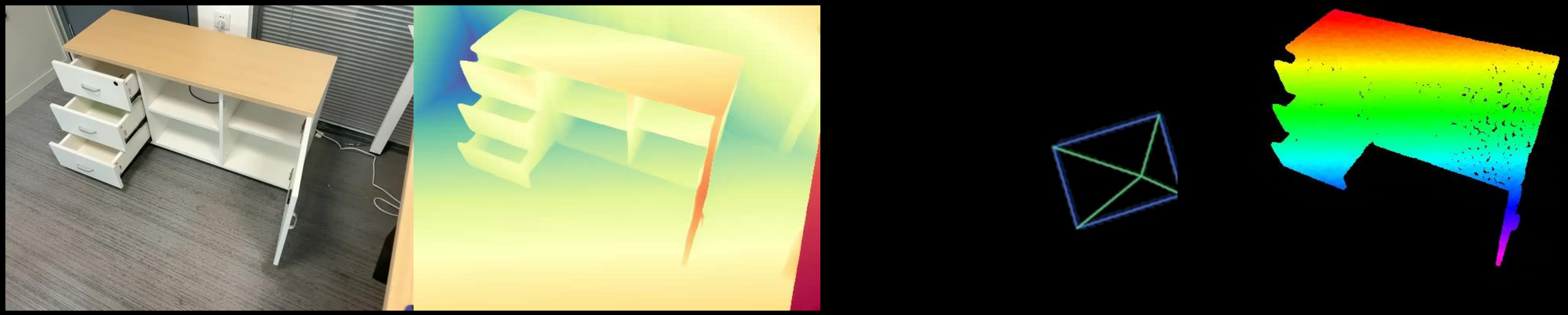
I_0 I_t

Camera, Depth, Tracks Estimation

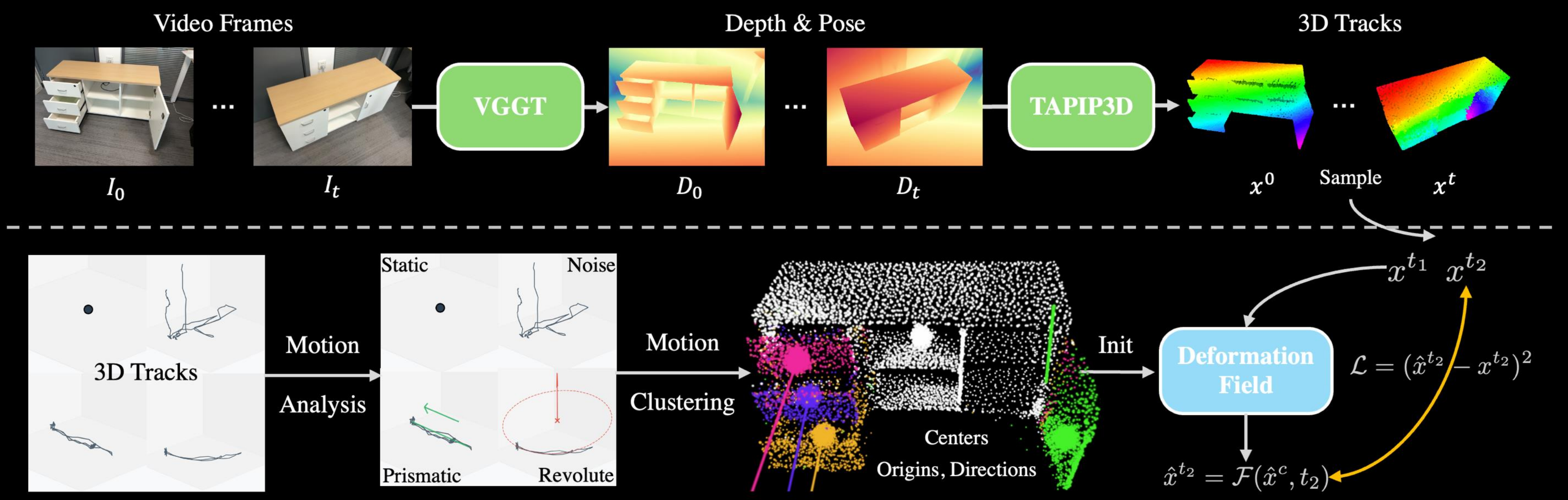


Camera, Depth, Tracks Estimation

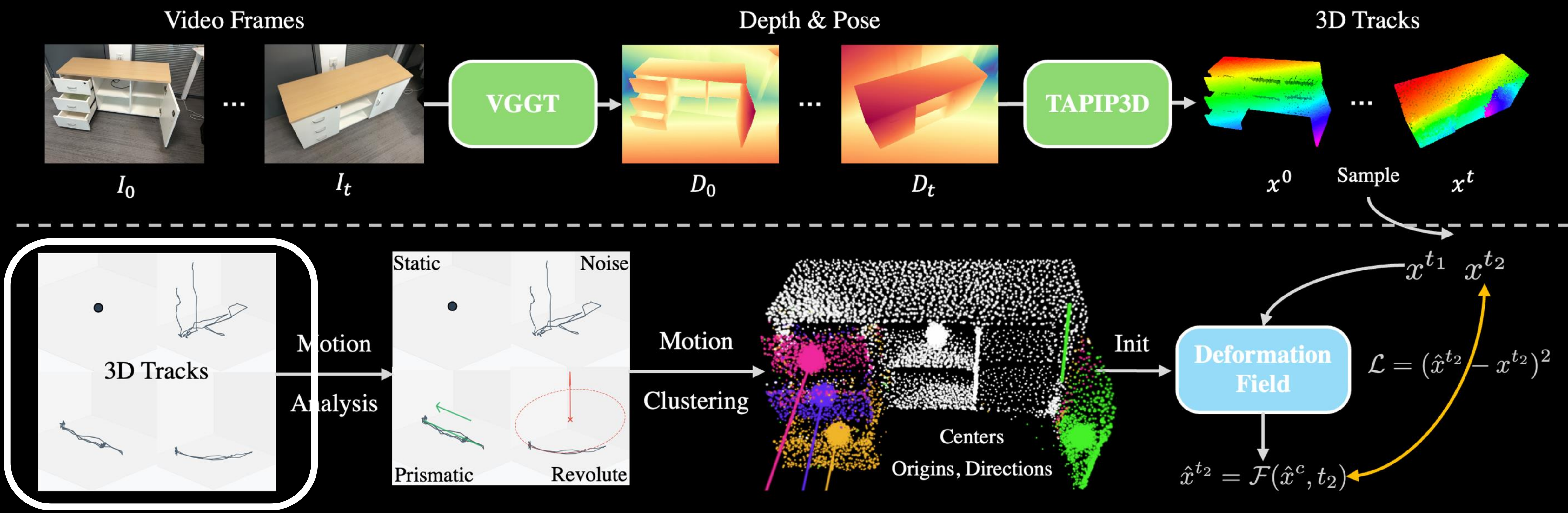




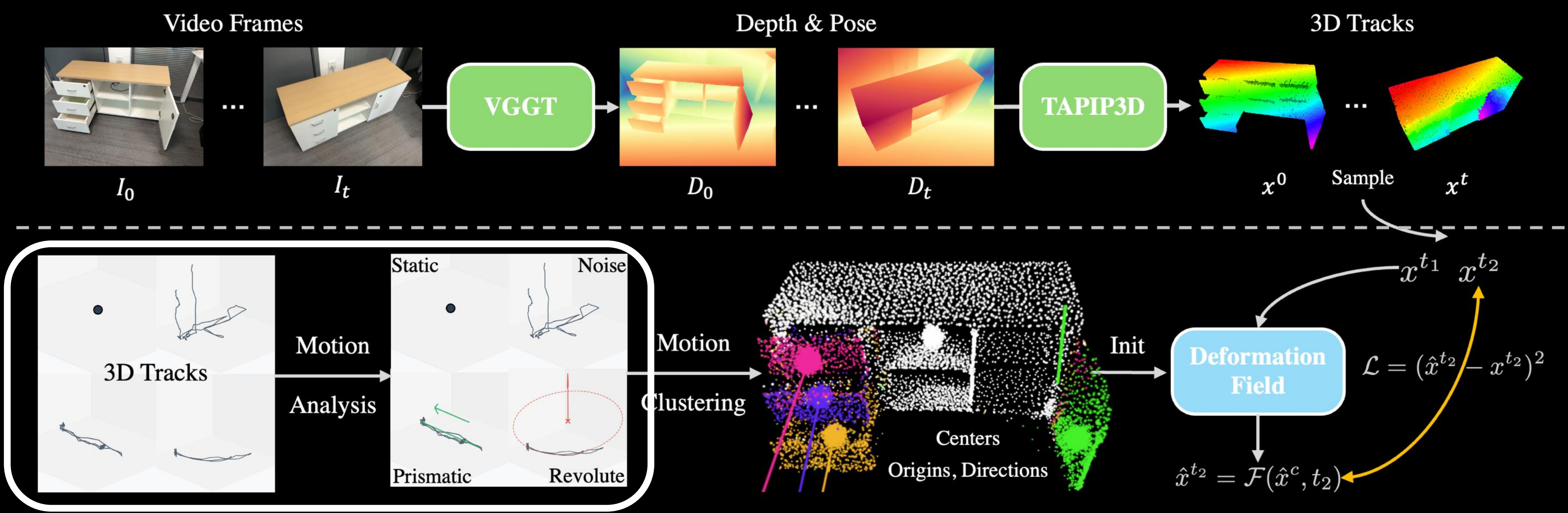
Motion Analysis & Deformation Field Initialization



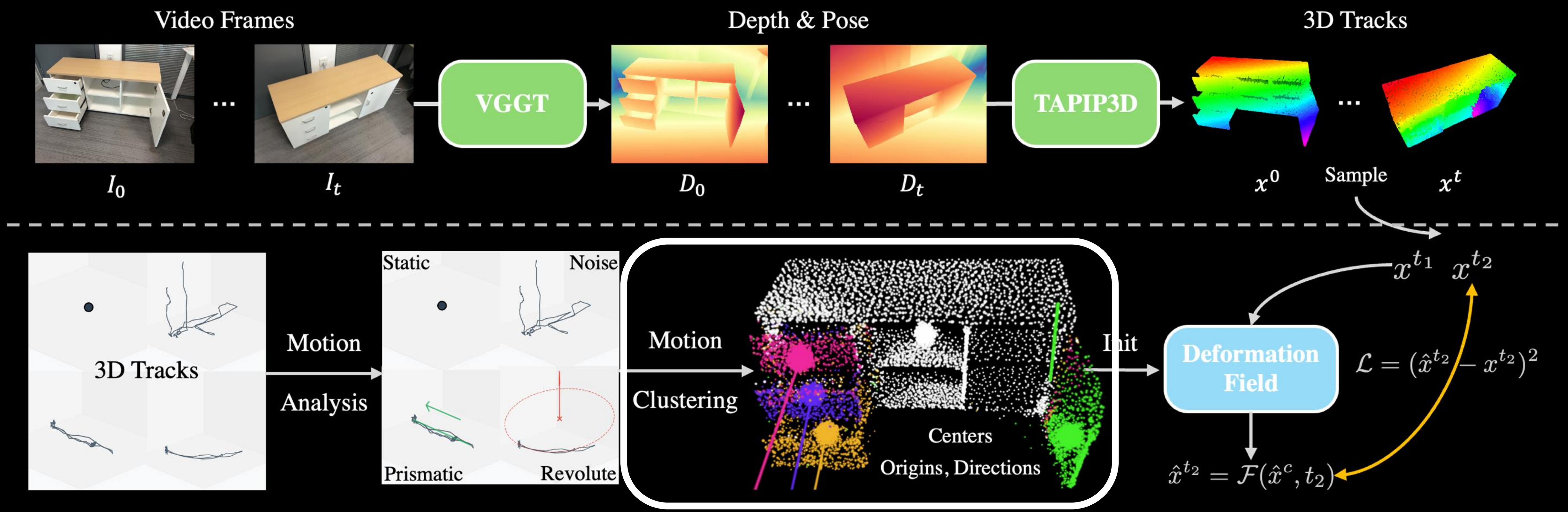
Motion Analysis & Deformation Field Initialization



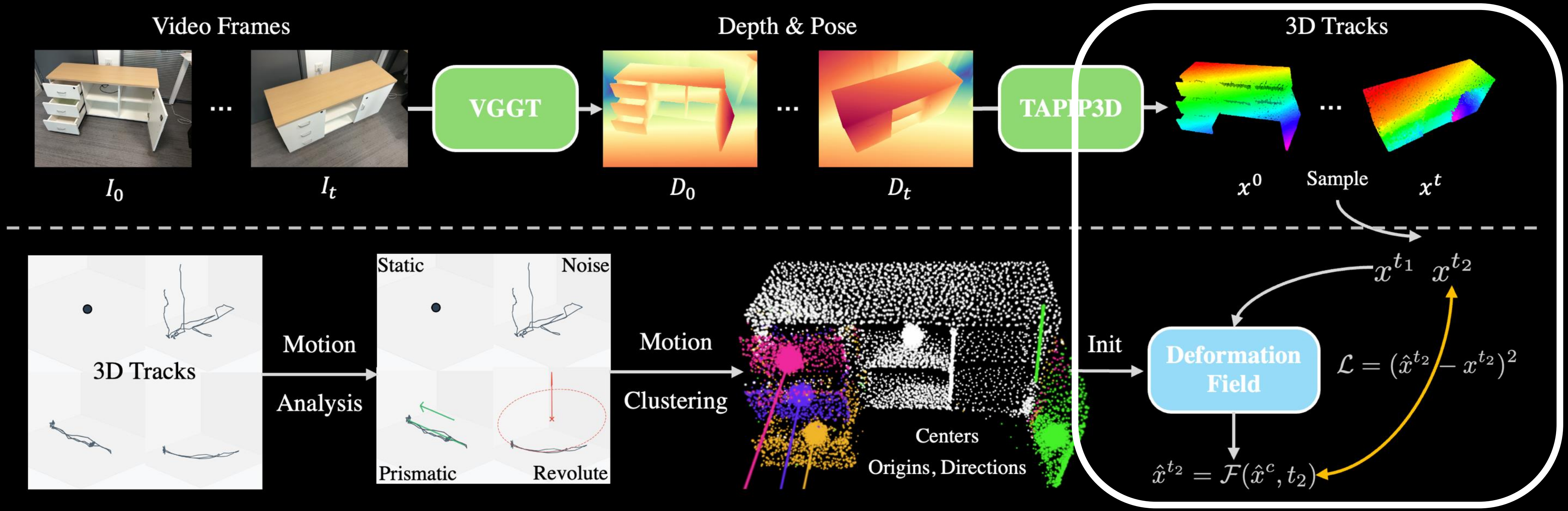
Motion Analysis & Deformation Field Initialization



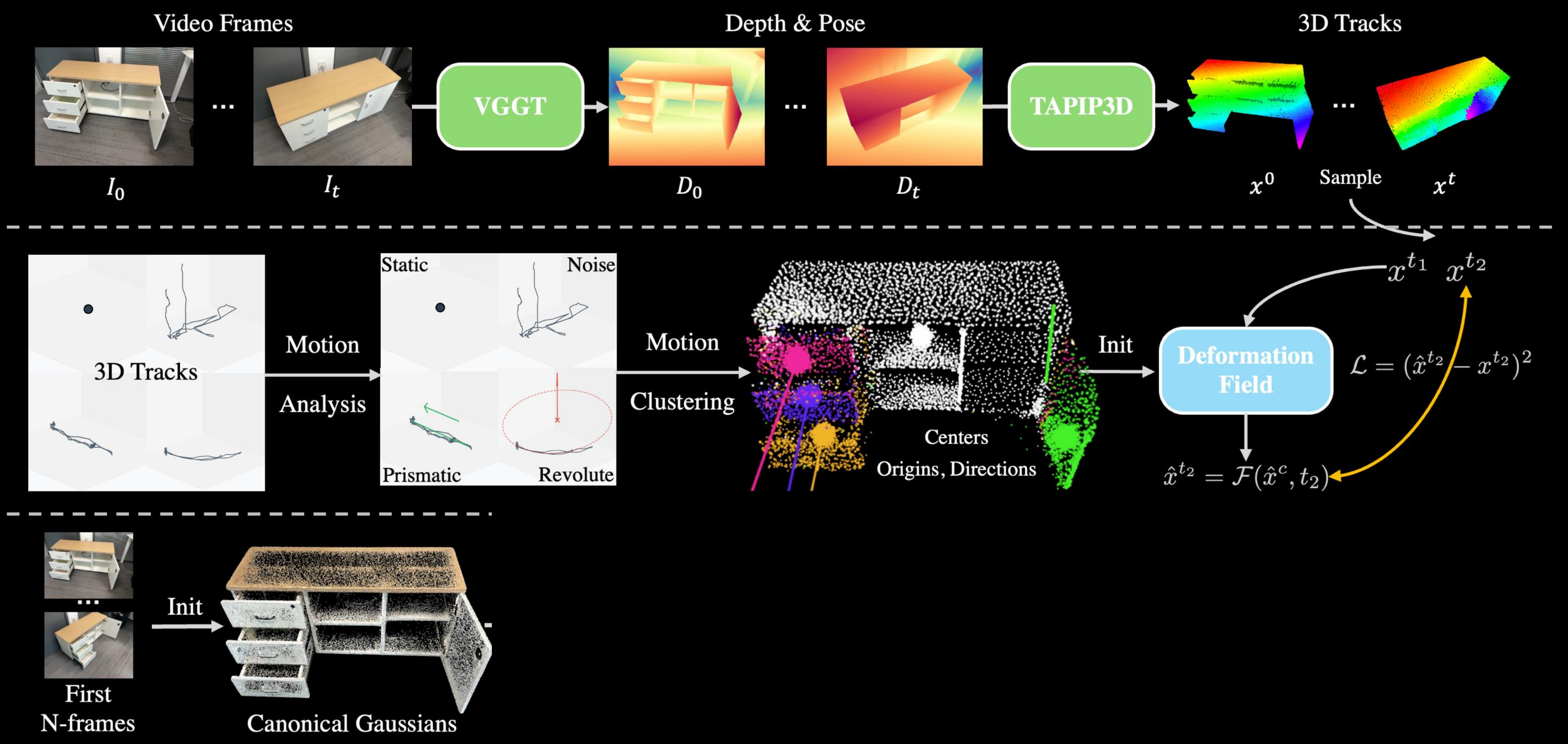
Motion Analysis & Deformation Field Initialization



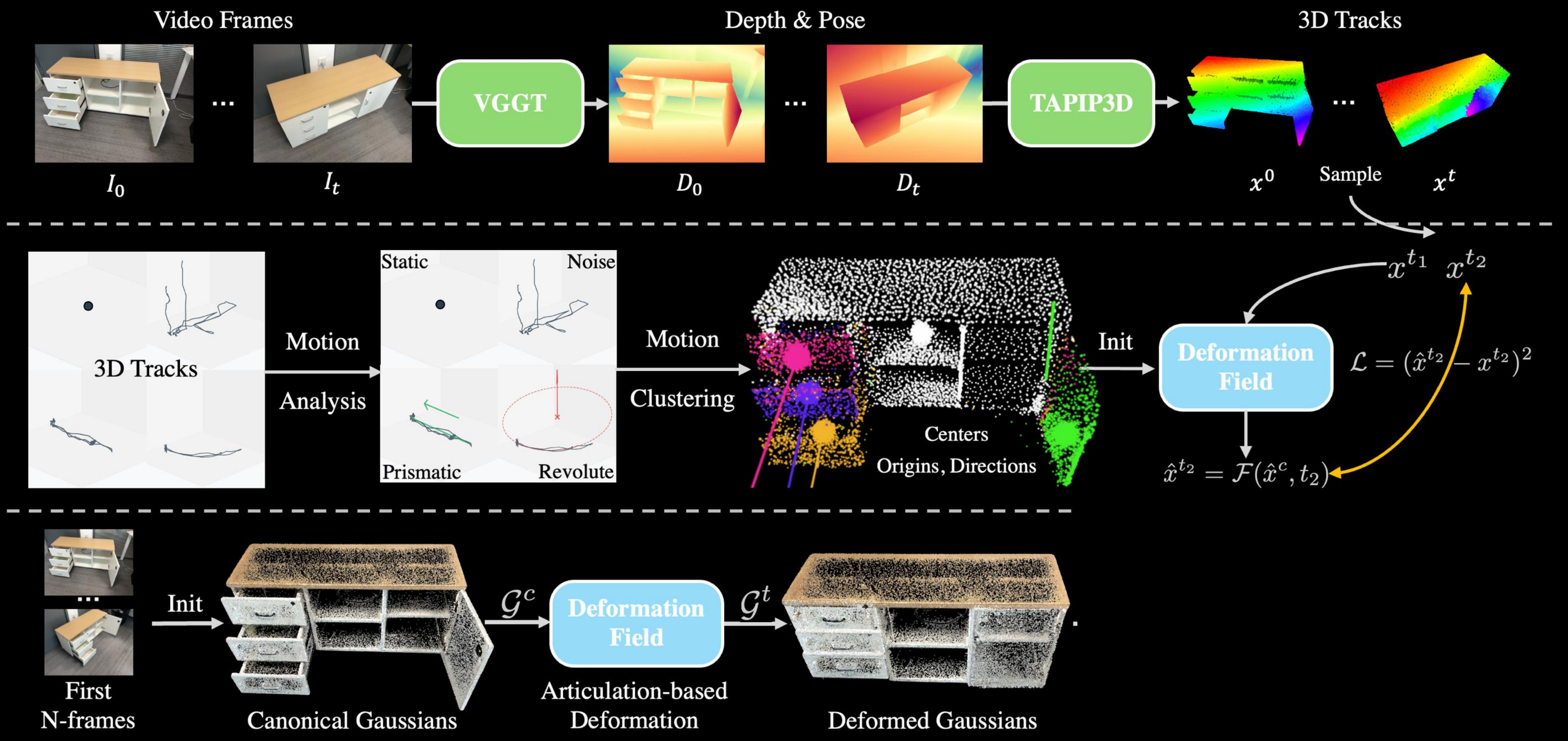
Motion Analysis & Deformation Field Initialization



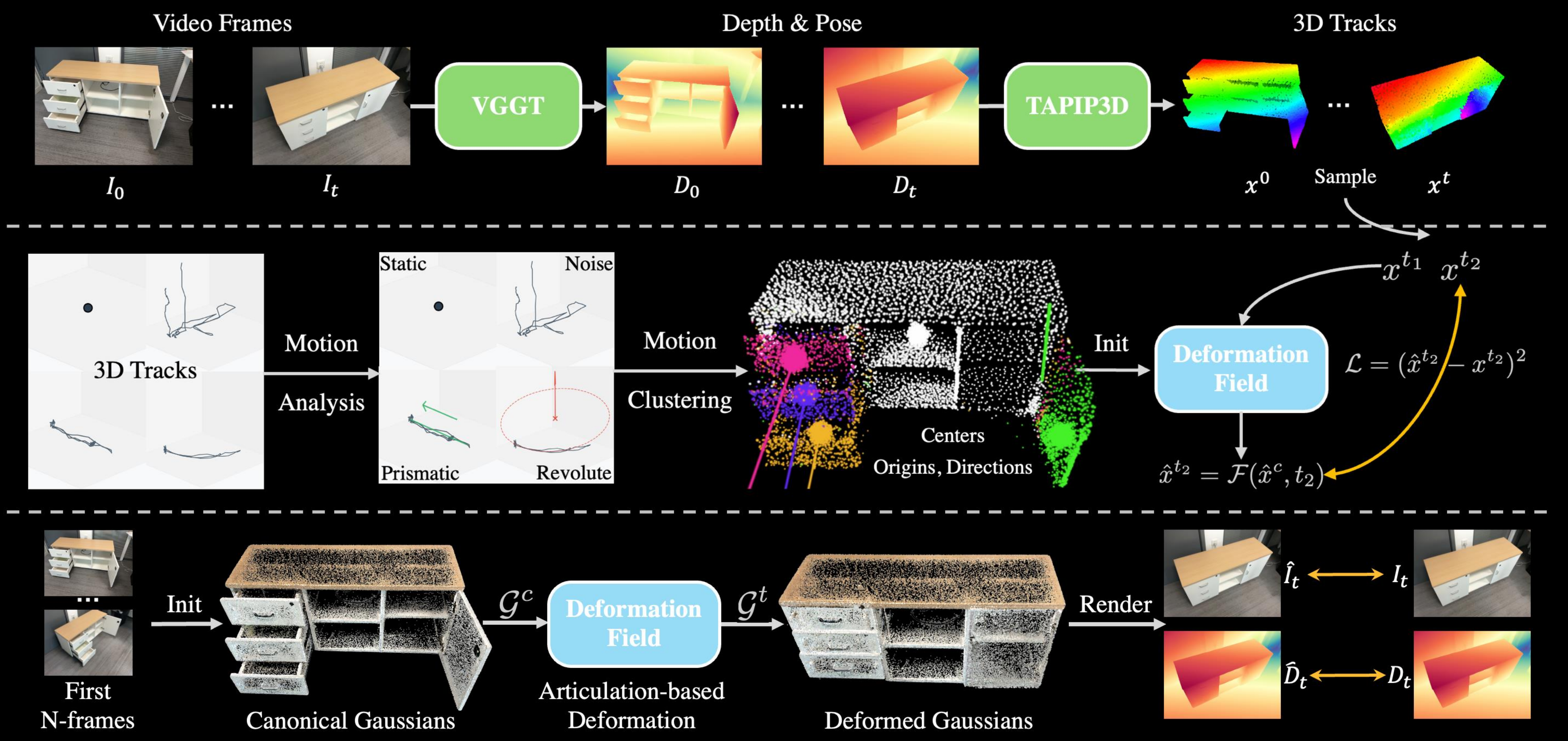
Geometry Reconstruction & Articulation Learning



Geometry Reconstruction & Articulation Learning



Geometry Reconstruction & Articulation Learning



Quantitative Comparison

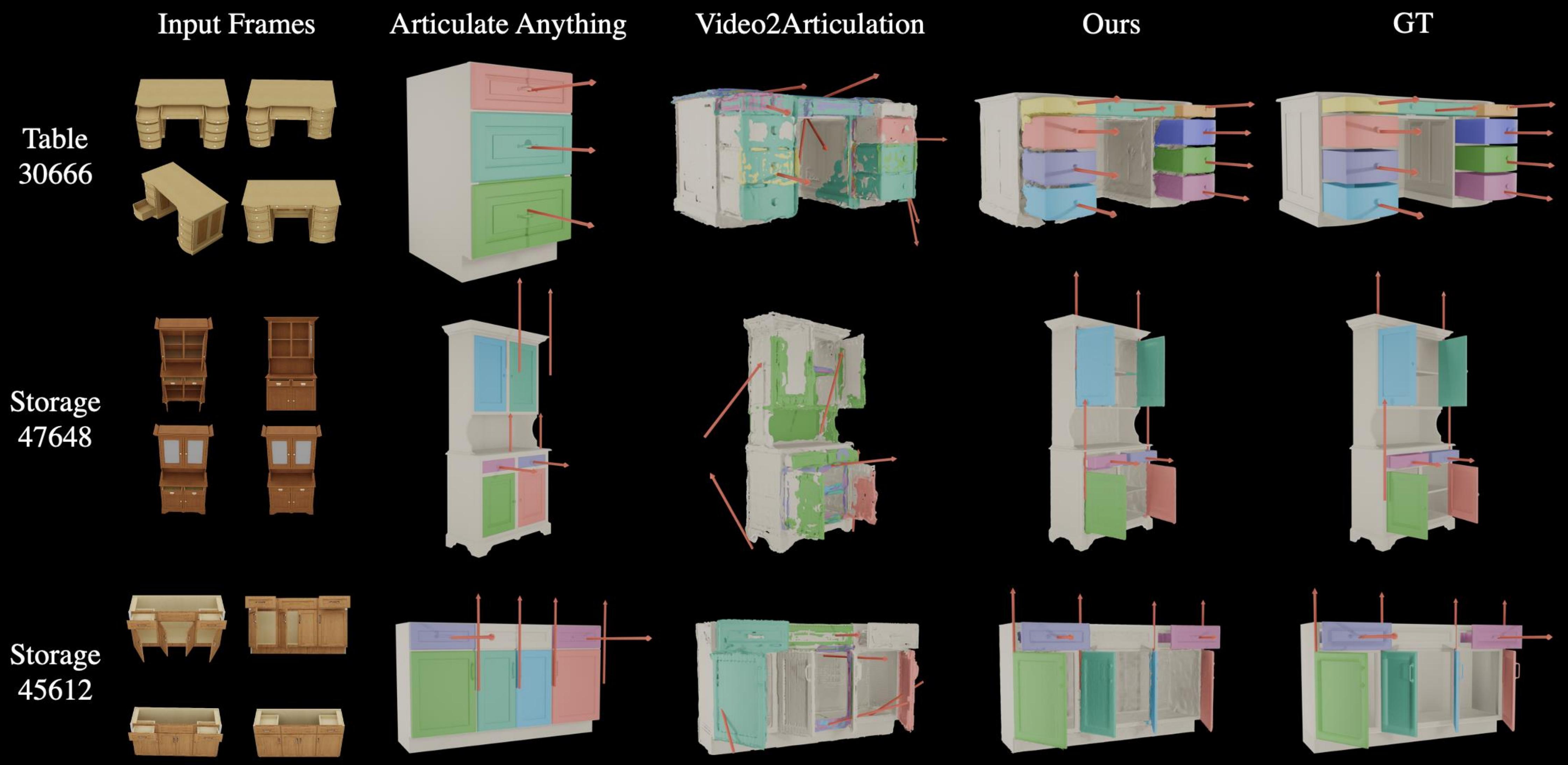
Method	Revolute Joint Estimation			Prismatic Joint Estimation		Reconstruction		
	Axis (○)	Position (cm)	State (○)	Axis (deg)	State (cm)	CD-w (cm)	CD-m (cm)	CD-s (cm)
ArticulateAnything† (Le et al., 2025)	46.98±45.27	81.00±40.00	N/A	52.71±44.69	N/A	11.00±22.00	59.00±73.00	7.00±18.00
RSRD† (Kerr et al., 2024)	67.06±29.22	203.00±748.00	59.02±34.38	69.91±24.07	70.00±48.00	339.00±2147.00	82.00±117.00	14.00±41.00
Video2Articulation† (Peng et al., 2025)	18.34±32.09	13.00±25.00	14.32±26.35	13.75±18.91	8.00±22.00	1.00±1.00	13.00±26.00	6.00±19.00
Video2Articulation (Peng et al., 2025)	13.83±28.15	11.55±22.39	10.25±21.27	14.37±19.08	3.44±6.25	3.45±16.46	12.21±24.44	5.39±17.09
Ours	**0.32±0.44**	**0.42±0.75**	**1.15±2.29**	**0.35±0.45**	**1.03±2.46**	**0.29±0.24**	**0.40±0.32**	**1.11±2.11**

Method	Axis (°)	Position(cm)	CD-w(cm)	CD-m(cm)	CD-s(cm)
ArticulateAnything (Le et al., 2025)	43.65 ± 44.72	15.66 ± 36.20	16.10 ± 37.34	17.66 ± 36.74	16.04 ± 37.36
Video2Articulation (Peng et al., 2025)	48.88 ± 24.18	37.04 ± 31.82	5.07 ± 21.78	30.63 ± 25.64	10.22 ± 22.23
Ours	**0.34±0.80**	**0.10±0.10**	**0.09±0.09**	**0.26±0.61**	**0.24±0.58**

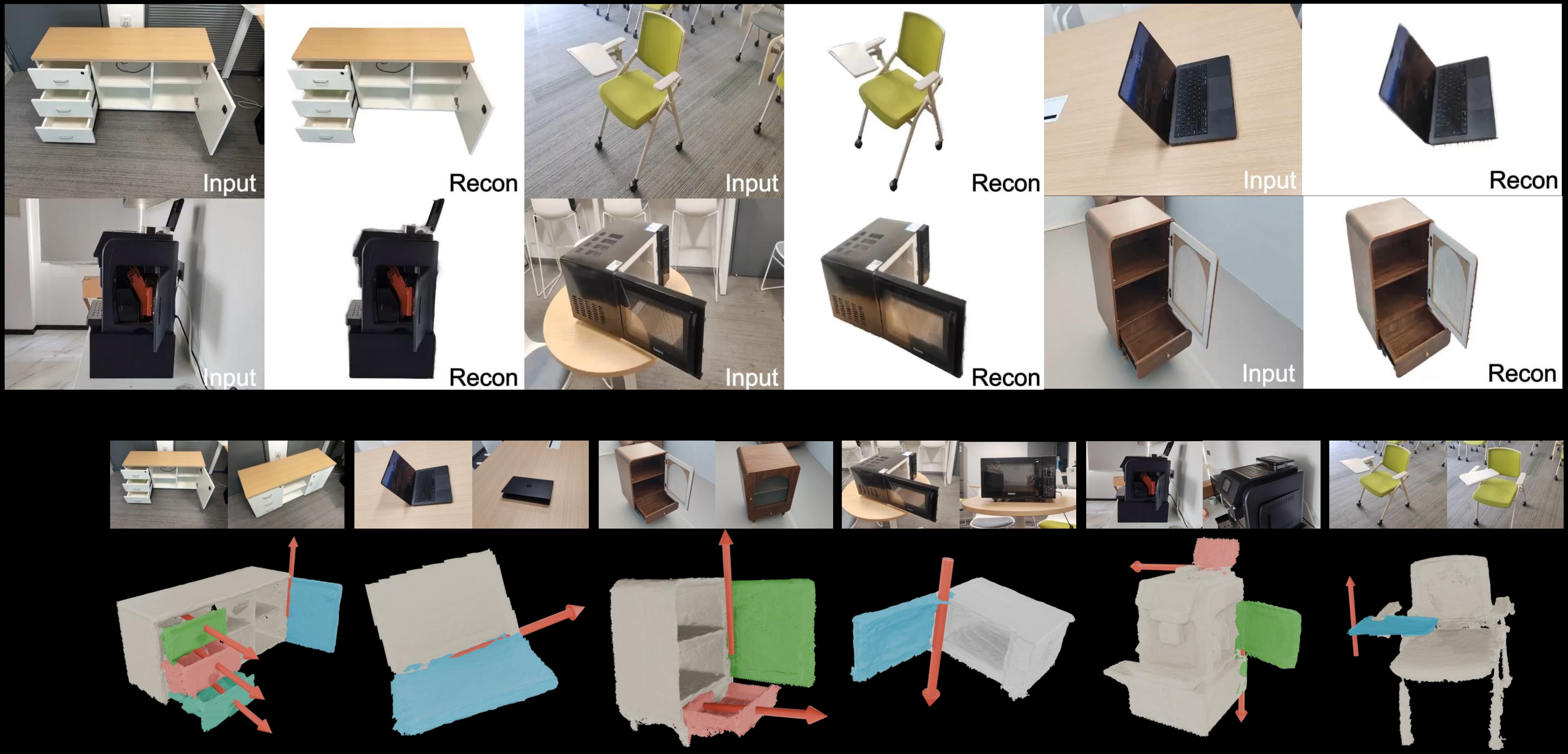
State-of-the-art performance on all metrics

Reducing the error by about two orders of magnitude

Qualitative Comparison



Real-world Experiments




Data Capture
Input
Recon
Gaussian

Takeaways

- Utilizing motion priors is crucial for dynamic object modeling
  - Articulated objects are still easy to model, priors or patterns are more difficult to be defined

Takeaways

- Utilizing motion priors is crucial for dynamic object modeling
  - Articulated objects are still easy to model, priors or patterns are more difficult to be defined
- Object-level articulated object reconstruction is do-able
  - Generating an interactable scene is still very difficult, due to both the increasing number of dynamic parts and occlusions

Takeaways

- ➢ Utilizing motion priors is crucial for dynamic object modeling
  - ❖ Articulated objects are still easy to model, priors or patterns are more difficult to be defined
- ➢ Object-level articulated object reconstruction is do-able
  - ❖ Generating an interactable scene is still very difficult, due to both the increasing number of dynamic parts and occlusions
- ➢ Monocular video with sufficient camera trajectory design gives good reconstruction results
  - ❖ How to utilize large-scale internet-scale egocentric interaction data remains a challenge

CLONE (CoRL 2025)

ControlVLA (CoRL 2025)

GWM: Towards Scalable Gaussian World Models for Robotic Manipulation

ICCV 2025

https://gaussian-world-model.github.io/

Thank you
Q&A

COLA (arXiv 2025)

SceneWeaver (NeurIPS 2025)

VideoArtGS: Building Digital Twins of Articulated Objects from Monocular Video

arXiv:2509.17647

https://videoartgs.github.io