VITA: Vision-to-Action Flow Matching Policy

Dechen Gao, Boqi Zhao, Andrew Lee, Ian Chuang, Hanchu Zhou, Hang Wang,
Zhe Zhao, Junshan Zhang, Iman Soltani
Website: *ucd-dare.github.io/VITA*

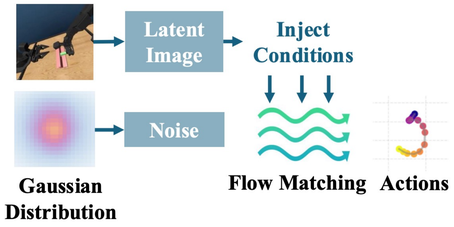
Website UC Davis UC Berkeley ICLR

Noise-Free and *Conditioning-Free* Flow Matching Policy for Fast/High-Precision Visuomotor Control

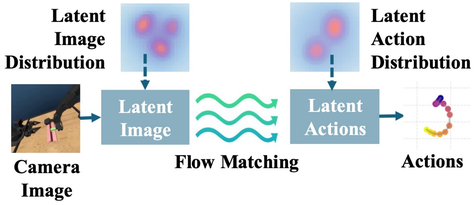
VITA flows from *latent visual representations* to *latent actions*

Flowing from Vision to Action

Conventional Flow Matching Policy



VITA Policy: *Noise-Free Flow Matching*



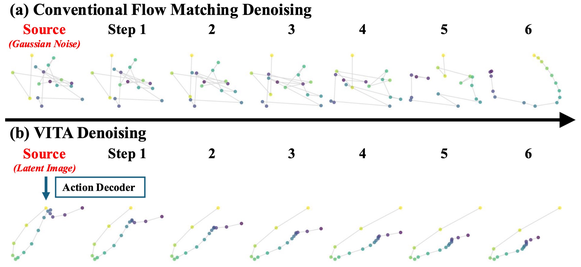
Conventional flow matching (FM) policies sample from Gaussian and *repeatedly* inject visual inputs via conditioning modules cross-attention during denoising, incurring time/space overheads.

VITA flows from visual to action latents. Because the source of the flow is visually grounded, VITA eliminates the need for conditioning modules and noise initialization

Aligning Vision and Action Manifolds

VITA learns ***action-centric visual representations***: visual latents can be decoded into smooth action trajectories and progressively refined by the FM ODE.

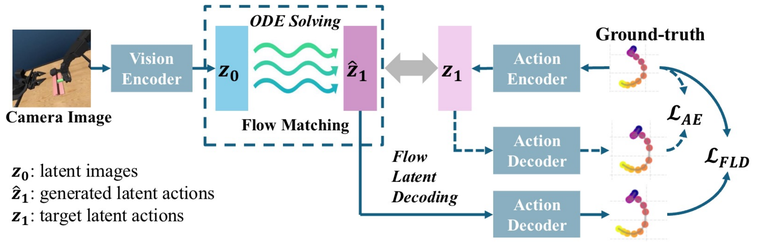
The ***closely aligned latent manifolds*** of vision and action reduces the complexity of the flow. We found that an ***MLP-only VITA*** flow network yields strong performance, whereas conventional FM necessitates complex transformers/U-Nets and MLP-only FM struggles due to the need to transport from unstructured Gaussian to structed actions.



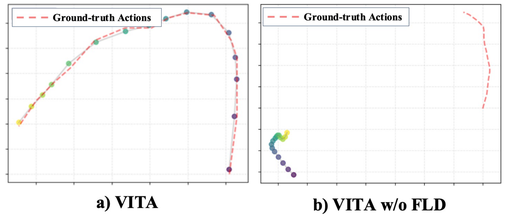
Conventional FM flows from uninformative noise to actions; while VITA learns action-centric visual representations, and flows from visually grounded sources.

End-to-End Latent Action Policy Learning

VITA is the first to end-to-end train latent action spaces with FM. We identify the main challenge for end-to-end (e2e) latent action FM policy: ***latent collapse*** caused by a training-inference gap. During training, the action decoder decodes encoder-based latents, whereas at inference, it uses ODE-generated latents.
We propose ***flow latent decoding*** (FLD) to bridge the gap by backpropagating through ODE solving steps anchor the ODE latent generation process using raw actions.



z_0: latent images
$\hat{z}_1$: generated latent actions
z_1: target latent actions



Right: Latent collapse during e2e latent action FM policy training. *Left:* we address latent collapse via flow latent decoding (FLD).

Simple & Efficient & High-Performance ***MLP-only VITA*** succeeds in real-world bimanual control!

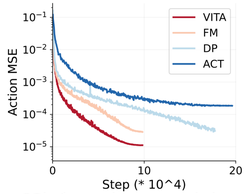
We use 9 simulated and 5 real-world tasks, covering bimanual manipulation with active vision (21D actions), and single-arm (2D, 7D, or 9D actions) tasks.

Task	VITA	FM	DP	ACT
ThreadNeedle	**91.33**±1.15	90±2	59.33±1.89	44.67±14.47
SlotInsertion	78±2	**82**±2	50.67±5.03	47.33±2.31
PourTestTube	78.67±2.31	**86**±2.31	46±0	42±7.21
HookPackage	**86**±2	82±2	37.33±6.11	32±2
CubeTransfer	**100**±0	**100**±0	94.67±3.06	99.33±1.16
PushT	**88**±2	83.33±1.16	74.67±6.11	28±5.29
Square	**95.33**±4.16	87.33±3.06	84±2	72±2
Can	**100**±0	**100**±0	95.33±1.16	88.67±2.31
CloseBox	**95.33**±1.16	85.33±2.31	85.33±1.16	72±5.29

VITA surpasses or matches SoTA methods in success rates.

Visual	Model	Architecture	Conditioning	Params	Latency	Memory
Vector	VITA	MLP	*N/A*	31.09M	**0.2215**	**333.86**
	FM	Transformer	AdaLN	31.16M	0.3307	410.38
	FM	U-Net	FiLM	84.05M	0.3650	818.79
	FM	MLP	AdaLN	32.20M	0.2831	413.95
	DDPM	U-Net	FiLM	81.82M	2.5985	801.47
Grid	VITA	Transformer	*N/A*	31.80M	**0.2502**	**377.55**
	FM	Transformer	Cross-Attn	29.06M	0.5102	529.16

VITA achieves 1.5x-2x speed up and 18%-28% memory reduction in real-time deployment.



High action precision and fast convergence.