GEVRM: Goal-Expressive Video Generation Model for Robust Visual Manipulation

Hongyin Zhang

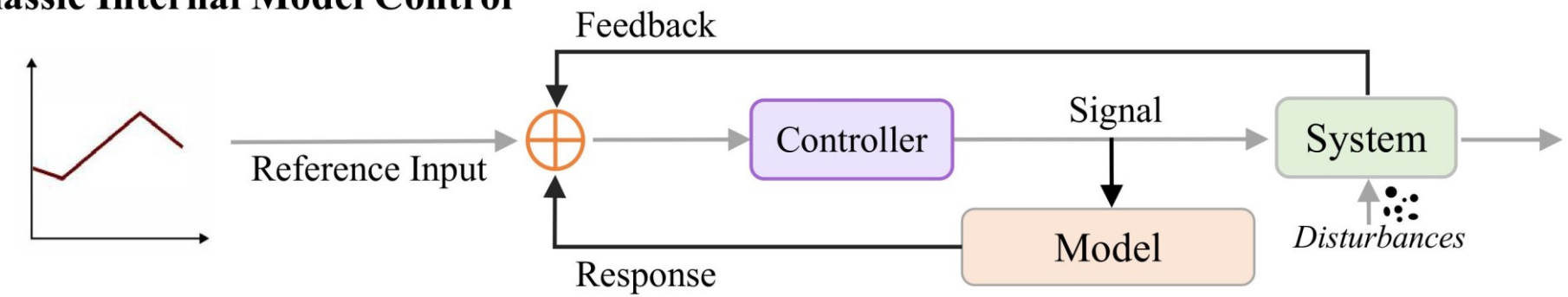
ICLR

西湖大學 WESTLAKE UNIVERSITY

机器智能实验室 PI: Donglin Wang
Machine Intelligence Laboratory (MiLAB)

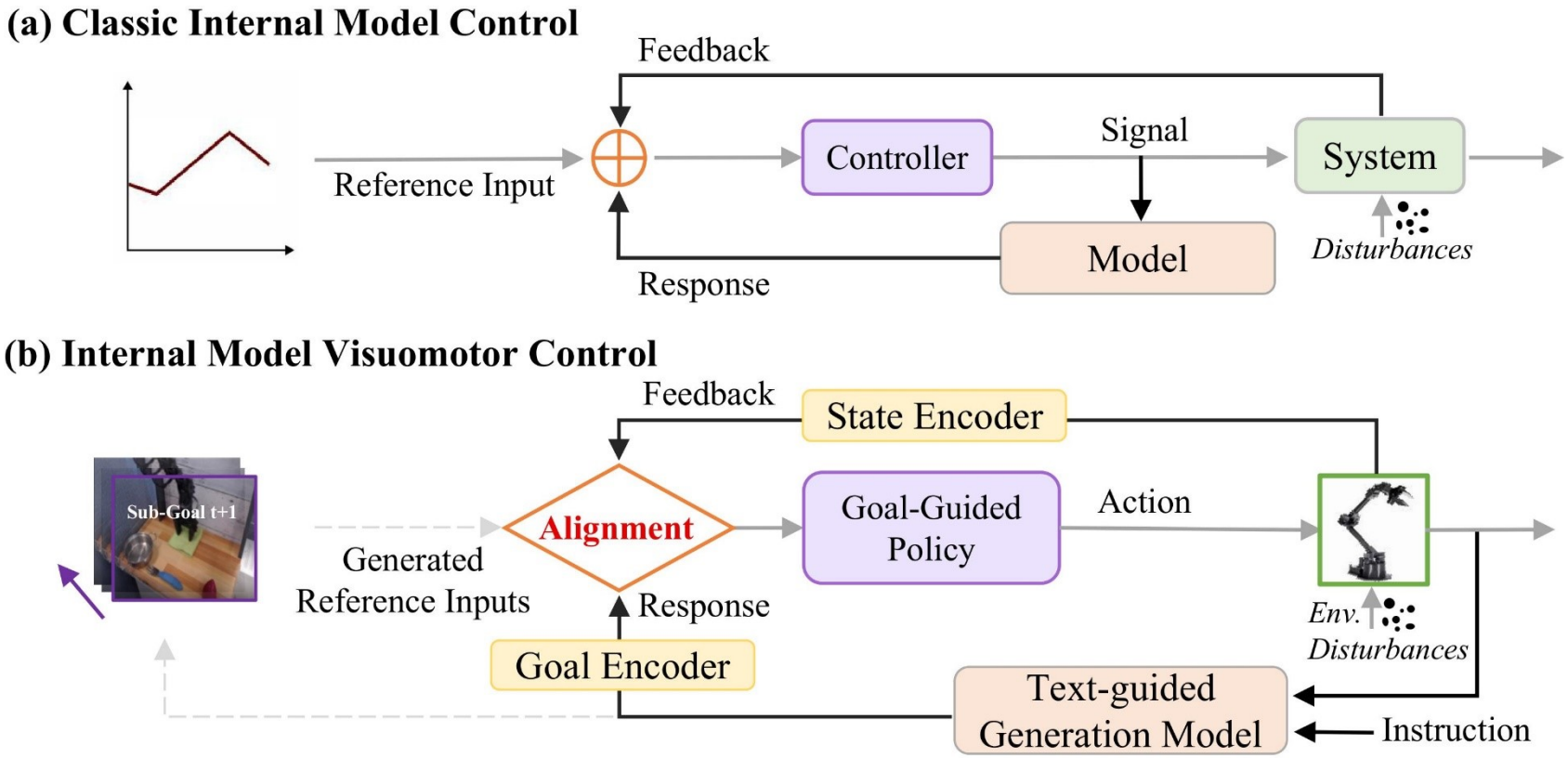
Motivation

(a) Classic Internal Model Control



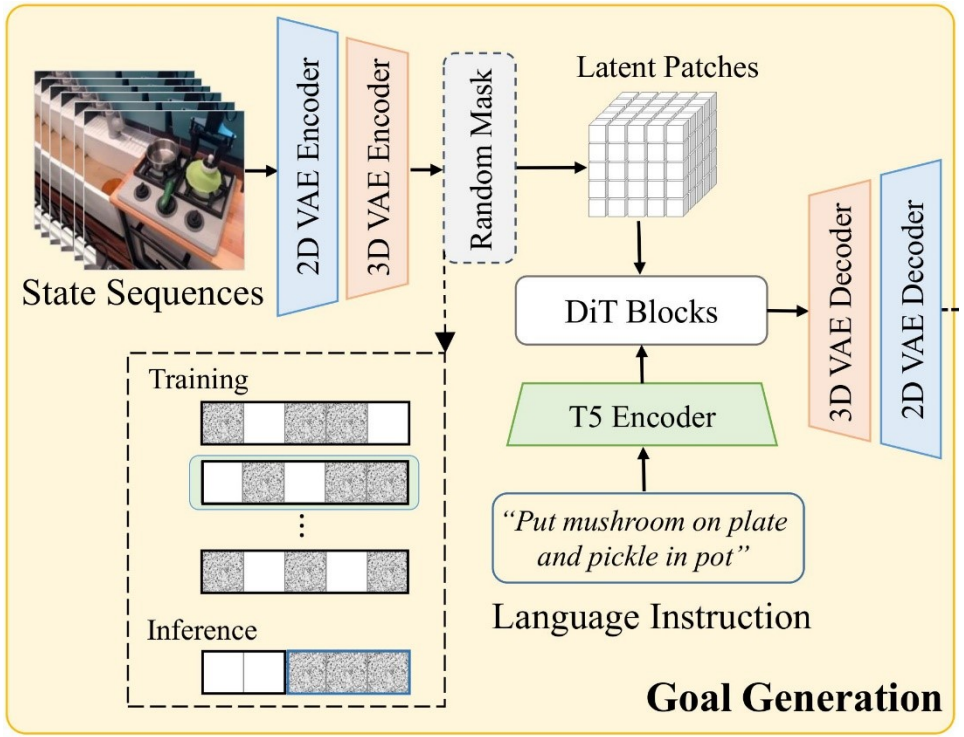
The classic **internal model control** (IMC) principle: A closed-loop system with an internal model that includes external input signals can accurately track the reference input and effectively offset the **disturbance**.

Motivation



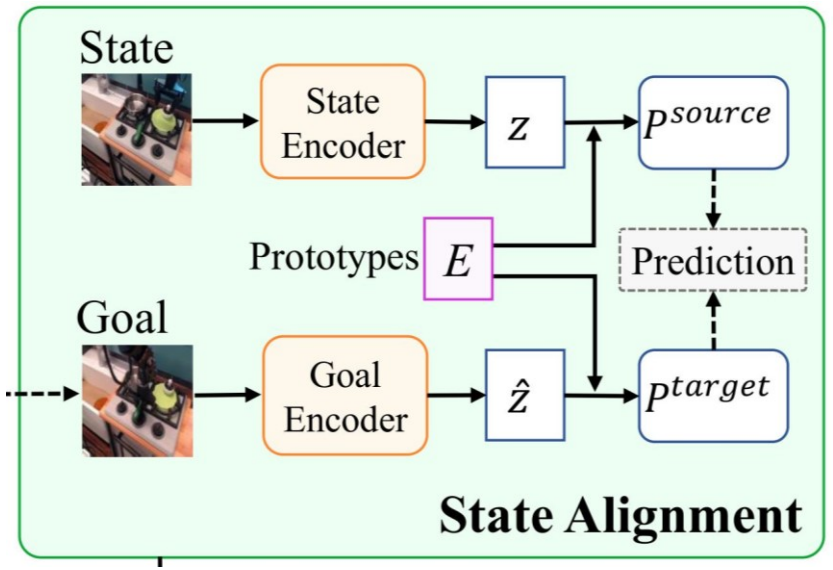
We propose **GEVRM**, a Goal-Expressive Video Generation Model for Robust Visual Manipulation. As shown in this figure, to effectively implement the classic IMC principle in the VLA model, some components of our method are adjusted accordingly.

Methodology



1. Robot behavior planner
   - Text-guided video diffusion model
   - Video spatiotemporal compression and random masking

Methodology



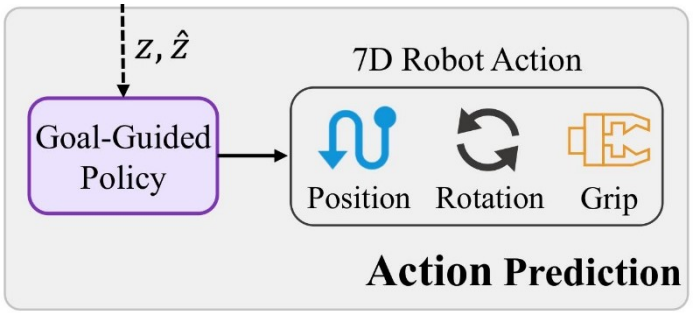
2. State alignment to simulate responses

$$z = \frac{f_\psi(x)}{\|f_\psi(x)\|_2}, \text{ and } \widehat{z} = \frac{f_{\psi'}(x_{goal})}{\|f_{\psi'}(x_{goal})\|_2}$$

$$p^{source} = \frac{e^{\frac{1}{\delta} z e_n}}{\sum_{n'} e^{\frac{1}{\delta} z e_{n'}}}, \text{ and } p^{target} = \frac{e^{\frac{1}{\delta} \widehat{z} e_n}}{\sum_{n'} e^{\frac{1}{\delta} \widehat{z} e_{n'}}}$$

$$\mathcal{J}_\psi = -\mathbb{E}_{x, x_{goal} \sim \mathcal{D}_{a,x}} (q^{source} \ln p^{target} + q^{target} \ln p^{source})$$

Methodology



3. Goal-guided policy

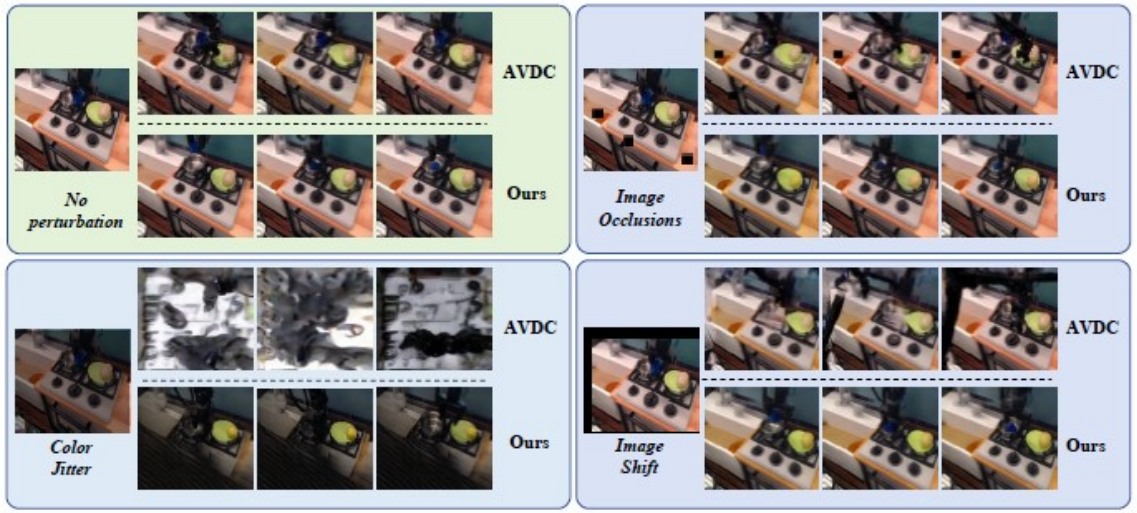
$$\mathcal{J}_\varphi = \mathbb{E}_{k\sim\mathbb{U}(1,K),\epsilon\sim\mathbb{N}(0,\mathcal{I}),x,x_{goal},a\sim\mathbb{D}_{x,a}}[\|\epsilon - \pi_\varphi(\sqrt{\widehat{\alpha}_k}a + \sqrt{1-\widehat{\alpha}_k}\epsilon), z, \widehat{z}, k\|_2]$$

$$a_{k-1} = \frac{1}{\sqrt{\alpha_k}}(a_k - \frac{\beta_t}{\sqrt{1-\widehat{\alpha}_k}}\pi_\varphi(a_k|z,\widehat{z},k)) + \sqrt{\beta_t}\epsilon, \text{ for } k = \{K,\cdots,1\}$$

4. Final optimization objective

$$\mathcal{J} = \mathcal{J}_\varphi + \lambda\mathcal{J}_\psi$$

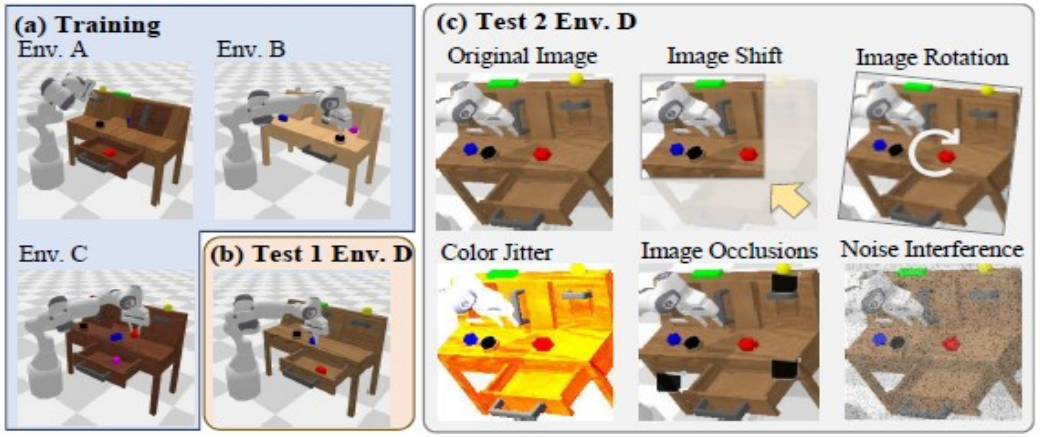
Goal generation quality comparison



Benchmark	Algorithms	FID (↓)	FVD (↓)	LPIPS (↓)	SSIM (↑)	PSNR (↑)
BridgeData	AVDC	246.45±39.08	22.89±4.99	0.23±0.03	0.73±0.05	18.22±2.53
BridgeData	SuSIE	114.79±21.38	–	0.22 ±0.08	0.71±0.07	16.39±2.90
BridgeData	GEVRM (Ours)	**35.70**±10.77	**4.16**±1.35	**0.06**±0.03	**0.89**±0.04	**22.36**±2.75
CALVIN	GR-1	236.75±38.87	12.83±2.60	0.20±0.02	0.65±0.03	18.59±0.95
CALVIN	SuSIE	214.14±45.45	–	0.15±0.04	0.75±0.05	18.12±2.29
CALVIN	GEVRM (Ours)	**94.47**±22.54	**3.80**±1.2	**0.09**±0.04	**0.80**±0.05	**21.10**±3.29

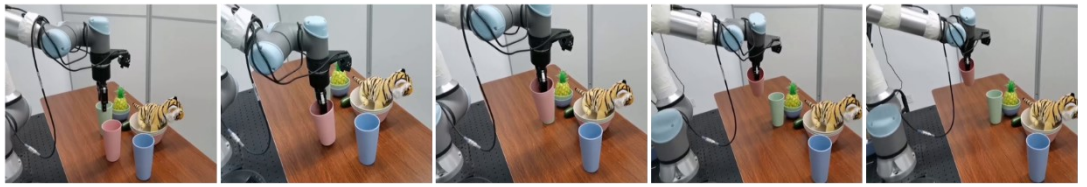
GEVRM can produce **fewer hallucinations** and generate **expressive goal states** following language instructions.

Action execution comparison

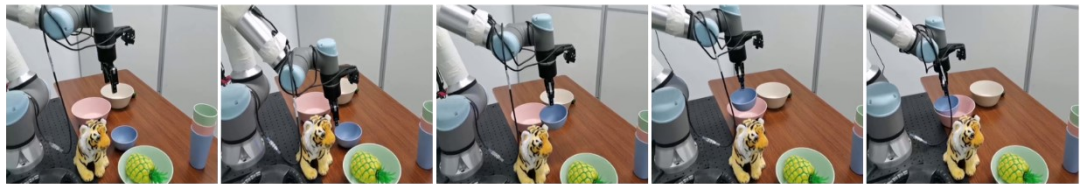


Five Perturbed Tasks	Algorithms	No. of Instructions Chained					Avg. Length (↑)
		1	2	3	4	5	
Average	SuSIE	0.56	0.26	0.13	0.10	0.06	1.11
	RoboFlamingo	0.63	0.35	0.18	0.09	0.05	1.31
	GR-1	0.67	0.38	0.22	0.11	0.06	1.44
	GEVRM (Ours)	**0.70**	**0.47**	**0.26**	**0.11**	**0.07**	**1.62**

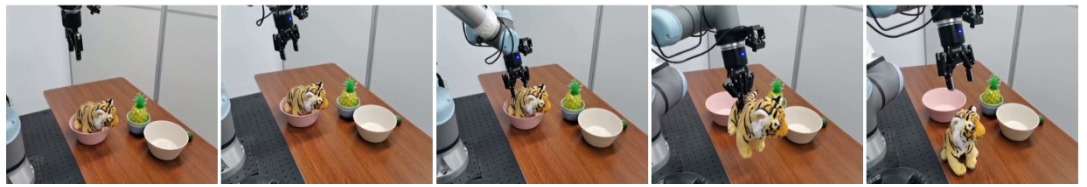
Instruction: Pick up the red cup. SR:0.8



Instruction: Put the smaller blue bowl into the red bowl. SR:0.8

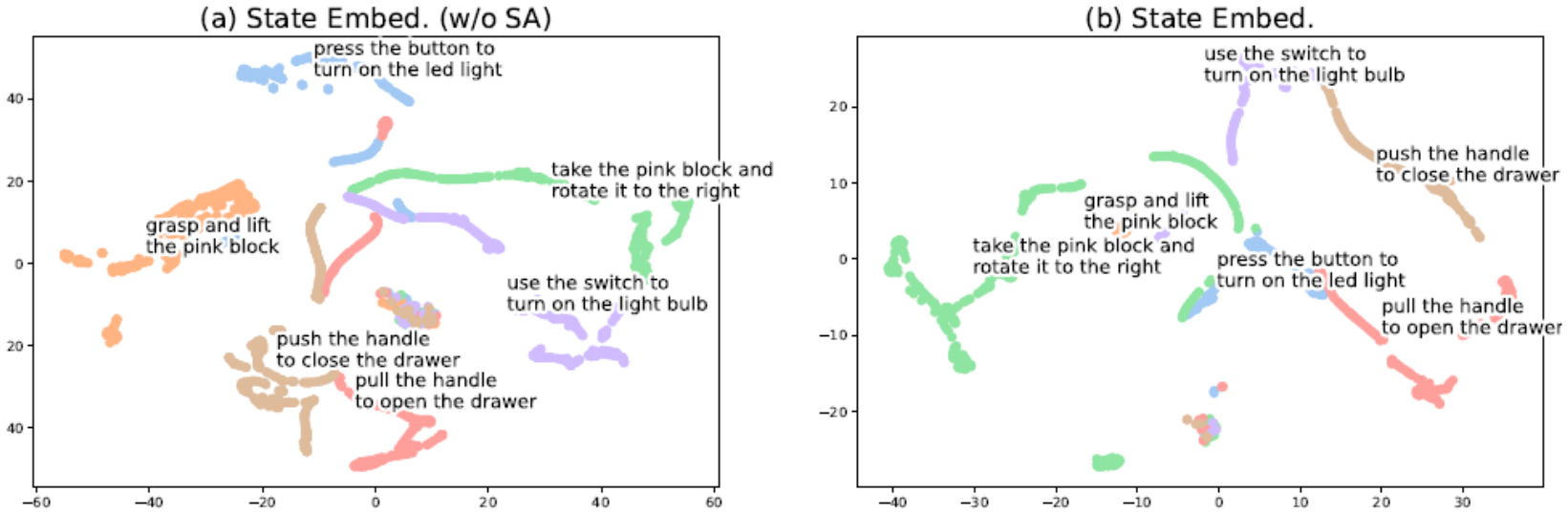


Instruction: Take the tiger out of the red bowl. SR:0.6



GEVRM can well simulate robot response and guide the policy to generate robust decision actions to resist external perturbations.

Ablation study



The representations with state alignment (SA) show enhanced cluster centers, class boundaries, and temporal consistency.

Summary

- We introduce GEVRM, a novel **robust** VLA model that incorporates the **IMC principle** to enhance robot visual manipulation.
- We study how to obtain **highly expressive goals** with a text-guided video generation model and **align state representations** through prototypical contrastive learning to resist external perturbations at deployment.
- Extensive experiments verify the effectiveness and advancement of the proposed GEVRM. It significantly outperforms the previous state-of-the-art on the CALVIN benchmark with standard and external perturbations. The expressiveness of the goal states generated in real visual manipulation is significantly improved compared to previous baseline methods.

Thank You!