

1 MOTIVATION

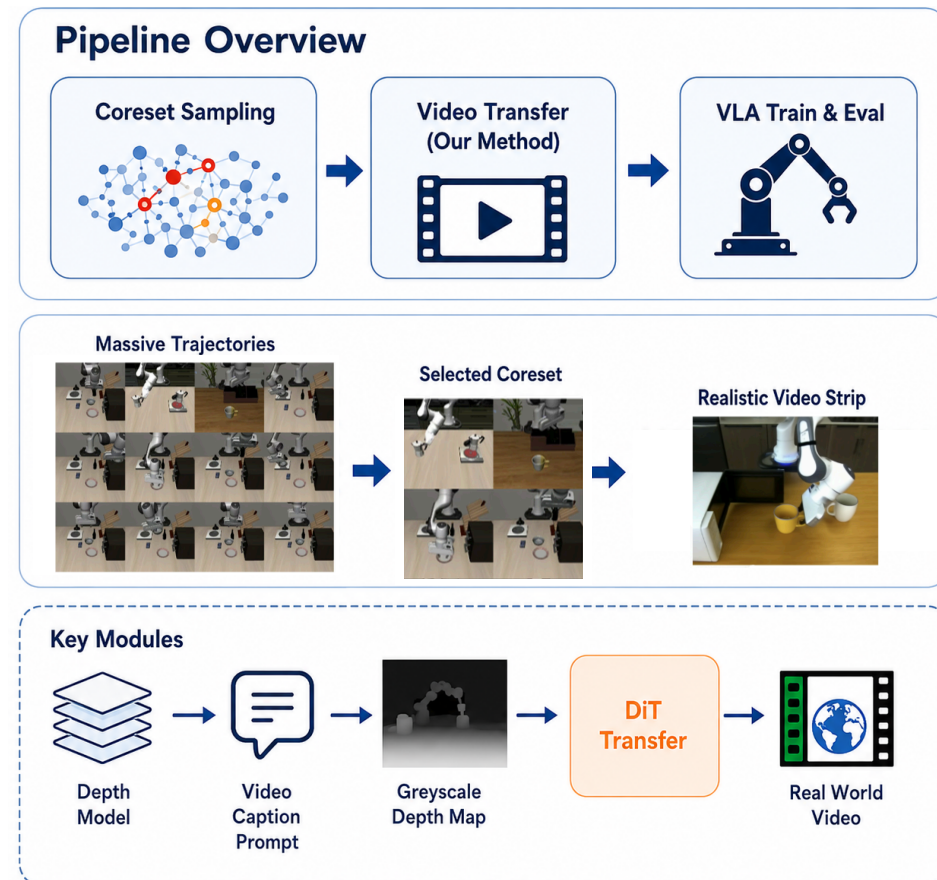
- Real-world data is **expensive** to collect
- Simulation is cheap but visually **unrealistic**
- Brittle generalization**: 95% → <30% under visual perturbations
- Root cause**: Models memorize, don't generalize

Can we bridge sim-to-real via video augmentation at scale?

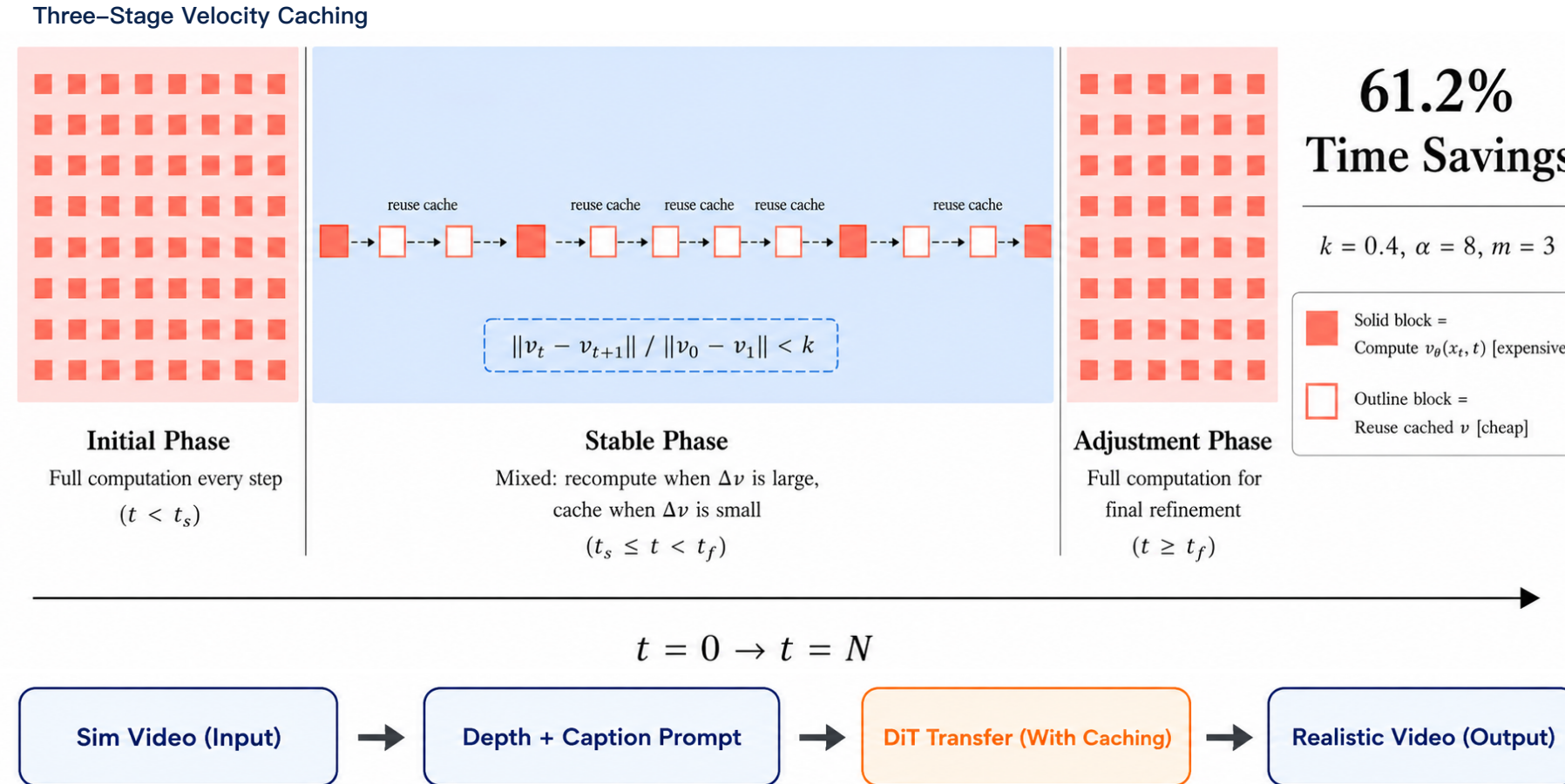
2 OUR CONTRIBUTIONS

- End-to-end sim-to-real video transfer** pipeline that converts simulation videos to realistic ones while preserving task semantics
- Velocity caching** exploits temporal redundancy in diffusion denoising → **60%+ faster** with negligible quality loss
- Coreset sampling** selects a small yet highly informative subset of trajectories by combining policy difficulty and visual diversity, enabling **effective augmentation at minimal cost**
- Consistent gains** across 4 benchmarks + real robot

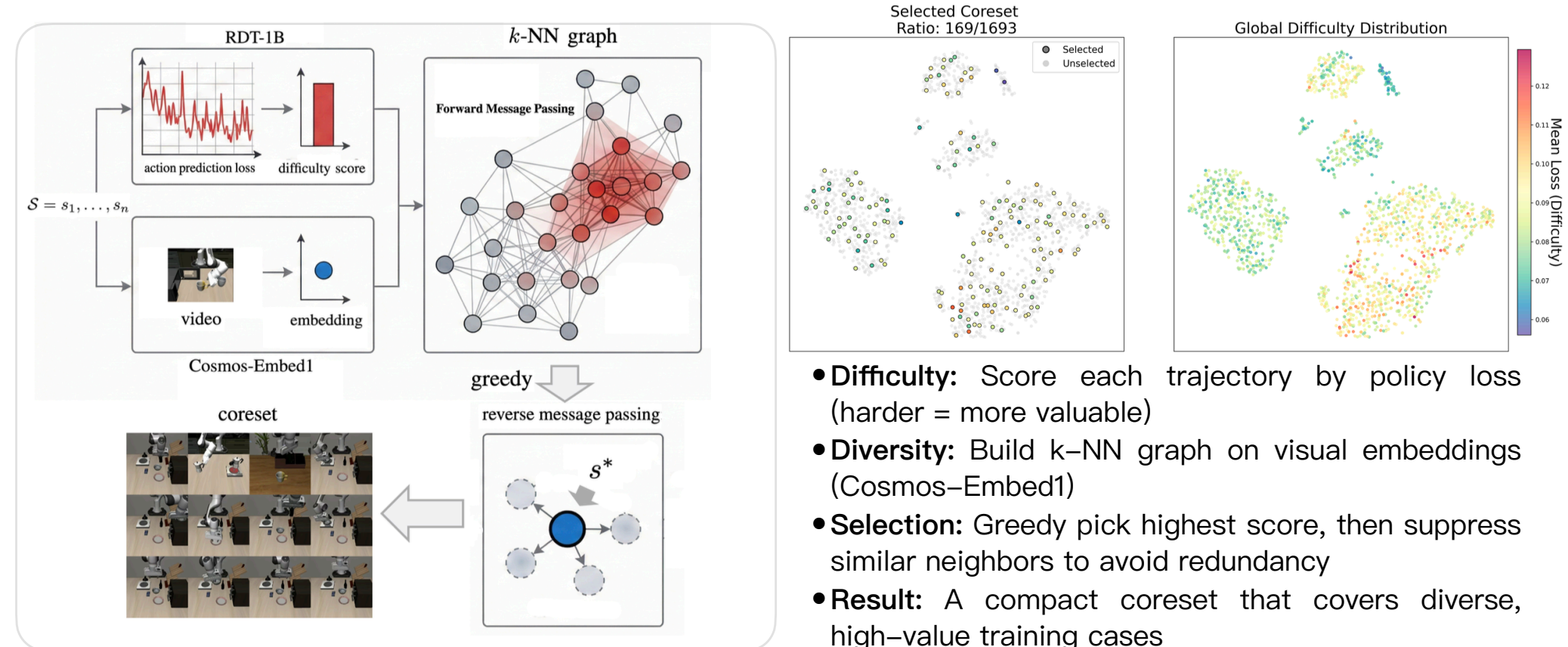
3 PIPELINE OVERVIEW



4 METHOD : EFFICIENT VIDEO TRANSFER



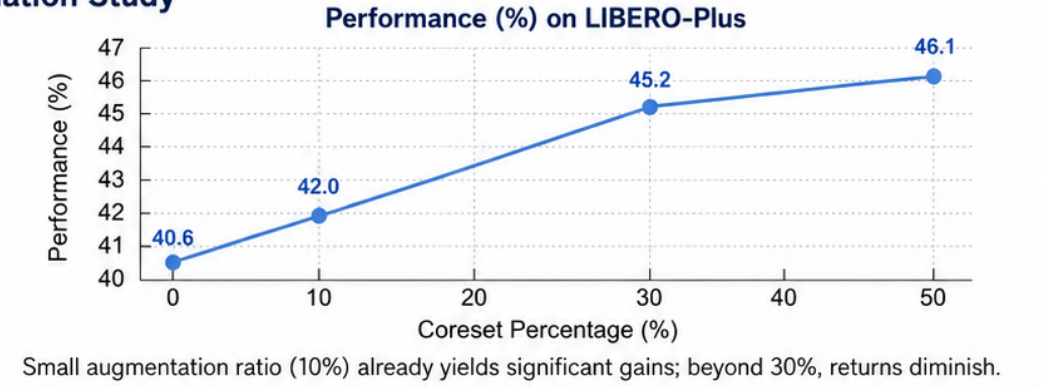
5 METHOD : CORESET SAMPLING



- Difficulty**: Score each trajectory by policy loss (harder = more valuable)
- Diversity**: Build k-NN graph on visual embeddings (Cosmos-Embed1)
- Selection**: Greedy pick highest score, then suppress similar neighbors to avoid redundancy
- Result**: A compact coreset that covers diverse, high-value training cases

5 MAIN RESULTS & ABLATION

1 Ablation Study



2 Main Results

Task Name	π_0			$\pi_{0.5}$		
	Ori.	Aug. (Rep)	Aug. (Mix)	Ori.	Aug. (Rep)	Aug. (Mix)
light conditions	75.0	75.7	78.7	94.5	97.9	99.3
objects layout	69.6	66.2	86.2	97.9	97.4	97.1
background textures	81.1	83.4	87.6	95.7	95.3	95.3
sensor noise	19.9	16.2	18.2	91.2	93.4	93.4
language instructions	37.9	43.2	55.9	90.0	90.0	89.1
robot initial states	10.3	8.1	6.3	82.6	84.9	83.2
camera view points	21.3	17.4	15.2	79.3	79.5	79.3
average	42.7	43.3	47.8	89.8	90.8	90.2

3 Real-world Results

Method	In-Dist.		Position		Background		Avg.
	Slot.	Stack.	Slot.	Stack.	Slot.	Stack.	
$\pi_{0.5}$	9/10	10/10	5/10	4/10	4/10	4/10	60%
+ Ours	10/10	10/10	7/10	6/10	6/10	5/10	73%
π_0	9/10	9/10	3/10	5/10	6/10	4/10	60%
+ Ours	10/10	10/10	5/10	8/10	5/10	7/10	75%

4 Video Transfer Quality

Task	Method	RMSE ↓	Abs.Rel ↓	Sq.Rel ↓	Sim ↑
adjust bottle	RoboTransfer	0.46	0.37	0.39	21.5
	Ours	0.28	0.16	0.07	26.3
beat hammer	RoboTransfer	0.30	0.28	0.16	22.3
	Ours	0.18	0.14	0.03	25.6
handover	RoboTransfer	0.49	0.32	0.20	18.6
	Ours	0.23	0.17	0.06	27.7
hanging mug	RoboTransfer	0.41	0.34	0.26	24.2
	Ours	0.13	0.12	0.02	26.4
pick bottles	RoboTransfer	0.37	0.30	0.18	20.3
	Ours	0.12	0.11	0.03	26.8

6 RESOURCES

Code



Paper

