IMAGINE THE FUTURE, INTERNALIZE THE GIST: EFFICIENT VLA REASONING VIA INTERNALIZED SPATIOTEMPORAL IMAGINATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision-language-action (VLA) models increasingly incorporate intermediate reasoning to improve robotic manipulation, yet existing approaches primarily reason about observed states without explicitly anticipating future scene evolution. Extending such reasoning to explicit future rollouts at every inference step, however, introduces substantial computational overhead. We propose **IG-VLA**, a VLA reasoning framework that enables models to ***i**magine the future and internalize the **g**ist*. Our **Latent Spatiotemporal Reasoning** learns to imagine task-relevant future scene evolution directly in visual representation space, guiding action prediction without costly pixel-level video generation. To further reduce inference overhead, we introduce **Scene Gist Memory**, which internalizes reasoning-derived scene–behavior associations into a compact Scene Gist Token, preserving the benefits of future reasoning while bypassing explicit future imagination at inference. Extensive experiments on LIBERO, LIBERO-Plus, and VLABench demonstrate the effectiveness and efficiency of IG-VLA. On the LIBERO-Plus Language suite, both the reasoning and gist policies outperform the strongest baseline by nearly 6% in success rate. The gist policy also achieves up to $6.38\times$ speedup over baselines, reducing inference latency from 1,081 ms to 169.5 ms per action chunk on a single NVIDIA A6000 GPU. These results demonstrate that future spatiotemporal reasoning can be effectively internalized for efficient VLA deployment.

1 INTRODUCTION

Vision-Language-Action (VLA) models enable general-purpose robotic manipulation by mapping multimodal observations and language instructions to actions Guan et al. (2025); Li et al. (2025a;b); Sapkota et al. (2025); Wang et al. (2025). However, learning an effective end-to-end mapping from multimodal perception to robotic actions requires extensive, high-quality demonstration data Liu et al. (2025a); Ma et al. (2026). To address this challenge, recent VLA models adopt a "think-before-acting" paradigm through Chain-of-Thought (CoT) reasoning Guo et al. (2025b); Ye et al. (2025); Yin et al. (2025).By reasoning over task goals, visual observations, and object relations before execution, these models decompose complex instructions and adapt action plans to environmental feedback.

Existing VLA paradigms include *language-based reasoning* Zawalski et al. (2024); Lee et al. (2025); Qu et al.; Huang et al. (2026a), which carry out reasoning in language domain and *reasoning in vision*, exemplified by CoT-VLA's decomposition of predicted goals into visual subgoals Zhao et al. (2025a) and VLA-Thinker's repeated incorporation of intermediate state images Wang et al. (2026). As shown in Figure 1(a), language-based reasoning loses fine-grained visual details and achieves lower success rates. In contrast, incorporating updated visual information during reasoning can mitigate drift from the actual environment Wang et al. (2026) (Figure 1(b)).

In general, existing approaches face two limitations: (i) reliance on observed states rather than future predictions, and (ii) substantial computational overhead from repeated reasoning rollouts. Specifically, existing approaches primarily reason over observed states and plan subsequent actions based on the current environment, without explicitly anticipating how the environment may evolve under different decisions. Meanwhile, reasoning in VLA models need to perform full spatiotemporal reasoning rollouts repeatedly at every inference step, introducing substantial computation overhead.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

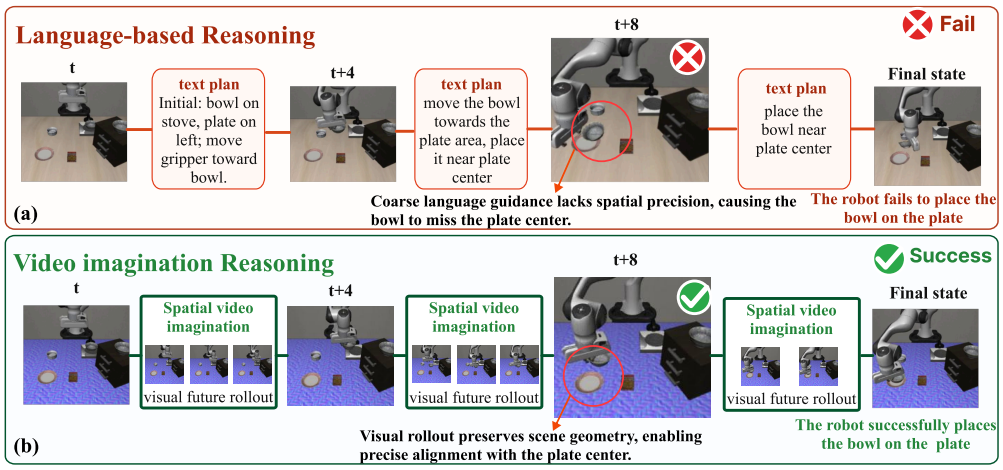

Figure 1: Language-based vs. latent visual reasoning. **(a)** Language-based reasoning may lose fine-grained spatial information, leading to imprecise actions; **(b)** Latent visual reasoning preserves evolving scene geometry through visual future reasoning, enabling more precise actions.

To address these issues, we propose a visual reasoning framework for VLA models with two core mechanisms: **latent spatiotemporal reasoning** and **scene gist memory**. Latent spatiotemporal reasoning enables the model to imagine future scene evolution in visual representation space, exploring plausible action trajectories and their potential outcomes. These imagined futures guide action selection, as shown in Figure 1(b). It is inspired by human reasoning process, which imagines how a scene might unfold under different candidate actions and makes decisions based on expected consequences Atance & O'Neill (2001). For example, when crossing a street, we consider multiple possible futures conditioned on our own actions such as crossing, waiting, or speeding up, then choose the safest action. This anticipation evaluates candidate behaviors through the imagined consequences rather than merely extrapolating scene evolution independently of actions Conti & Irish (2021).

To reduce computational overhead, our proposed Scene Gist Memory internalizes reasoning-derived scene–behavior associations into a compact *Scene Gist Token*. This token captures general scene context and corresponding action patterns rather than preserving the complete future reasoning trajectory. During inference, the gist encoder predicts the Scene Gist directly from the current context, allowing the model to bypass the explicit future-reasoning pathway while retaining reasoning-derived behavioral guidance (Figure 2). It's inspired by human behavioral psychology that repeated experience in familiar environments helps consolidate learned behaviors into procedural memory involving the basal ganglia Wulf et al. (2010); Wood & Rünger (2016). This memory supports rapid responses with less deliberation, reducing cognitive effort and decision time. For example, familiarity with a home allows a person to head directly to the kitchen for coffee without searching the entire house. Repeated experience reinforces the association between the familiar scene and the corresponding navigation behavior. Our main contributions are summarized as follows:

- Our framework incorporates **Latent Spatiotemporal Reasoning**, enabling VLA models to learn to imagine task-relevant scene evolution in visual representation space. These anticipated scene dynamics guide action prediction without costly pixel-level video generation.

- We further introduce **Scene Gist Memory**, which internalizes reasoning-derived scene–behavior associations into a compact Scene Gist Token. It enables the model to reuse this knowledge in corresponding scenarios, preserving reasoning benefits while bypassing explicit future imagination and full reasoning rollouts at inference.

- Extensive experiments on LIBERO, LIBERO-Plus, and VLABench demonstrate the effectiveness and efficiency of our framework. For example, IG-VLA achieves an 85.6% success rate on LIBERO-Plus's challenging Language suite, outperforming the strongest baseline by nearly 6%. Meanwhile, the gist policy requires only 169.5 ms per action chunk on a single NVIDIA A6000, achieving 6.38×, 1.78× and 1.55× speedups over FastWAM-IDM (1,081 ms), FastWAM (1-step) (302 ms) and ImageWAM (263 ms), respectively.

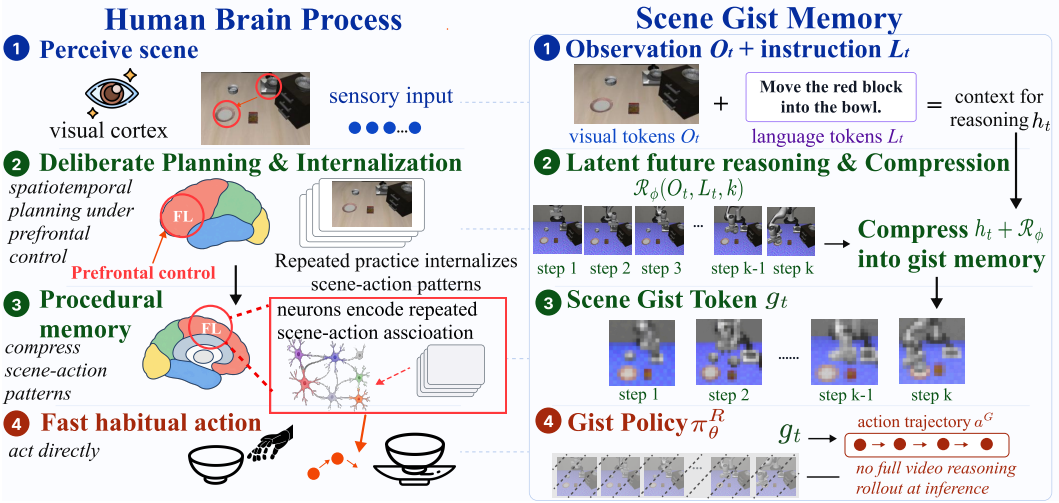

Figure 2: Human Automaticity vs. Scene Gist Memory.

2 Related Work

VLA Models. Recent advances in embodied AI have shifted robot learning toward foundation VLA models that directly map visual observations and language instructions into executable actions Brohan et al. (2022); Mees et al. (2024); Nvidia et al. (2025); Yang et al. (2025); Pertsch et al. (2025); Driess et al. (2026). OpenVLA Kim et al. (2024) adopts an end-to-end policy learning framework for language-conditioned robotic manipulation. π_0 Black et al. (2024) extends this paradigm with scalable policy learning across more diverse robotic tasks. HAMSTER Li et al. (2025c) strengthens spatial reasoning through structured visual representations, while TraceVLA Zheng et al. (2025) further incorporates 2D visual trajectories to enhance the alignment between perception and action.

VLA Reasonings. CoT improves LLM performance in mathematical reasoning, code generation, and question answering Wei et al. (2022); Yeo et al. (2025), motivating explicit reasoning in VLA models Huang et al. (2026b). Existing approaches mainly follow several directions. Prompting-based methods such as EcoT Zawalski et al. (2024) use supervised reasoning traces to decompose tasks into intermediate subgoals before action execution. Data-centric methods, including RAD Clark et al. (2025), combine human videos with LLMs to construct reasoning-action demonstrations for policy learning. Another line of work, represented by CoT-VLA Zhao et al. (2025b), replaces textual reasoning with visual subgoal representations to better align planning with robot perception. Recently, reinforcement learning Shao et al. (2024); Guo et al. (2025a) has been adopted to optimize reasoning policies under verifiable rewards, where GRPO encourages more effective reasoning trajectories Feng et al. (2026); Tan et al. (2025); Liu et al. (2025b). However, current methods often require costly reasoning supervision and sophisticated reward engineering Huang et al. (2026a).

3 Methodology

In this section, we present IG-VLA, a VLA framework that incorporates two core mechanisms, reasoning over possible future predicted trajectories to improve robustness and task success in long-horizon embodied tasks, and scene gist token to capture a coarse-grained scene-behavioral impression for efficient inference, as shown in Figure 3.

3.1 Model Architecture

We first introduce our overall model architecture, as illustrated in Figure 3. Our model builds upon a conventional VLA architecture, and incorporates spatiotemporal reasoning modules and gist modules.

❶ **Vision Encoder and Vision-Language Backbone.** A pretrained vision encoder E_v maps the current visual observation o_t to visual tokens: $e_t = E_v(o_t)$. These tokens compactly encode scene semantics and spatial structure, supporting subsequent spatiotemporal reasoning. A pretrained vision-language backbone $\mathcal{V}$ jointly processes e_t and the tokenized language instruction L_t: $h_t = \mathcal{V}(e_t, L_t)$, where h_t encodes the current observation and task instruction, providing multimodal context for future-scene reasoning and action generation.

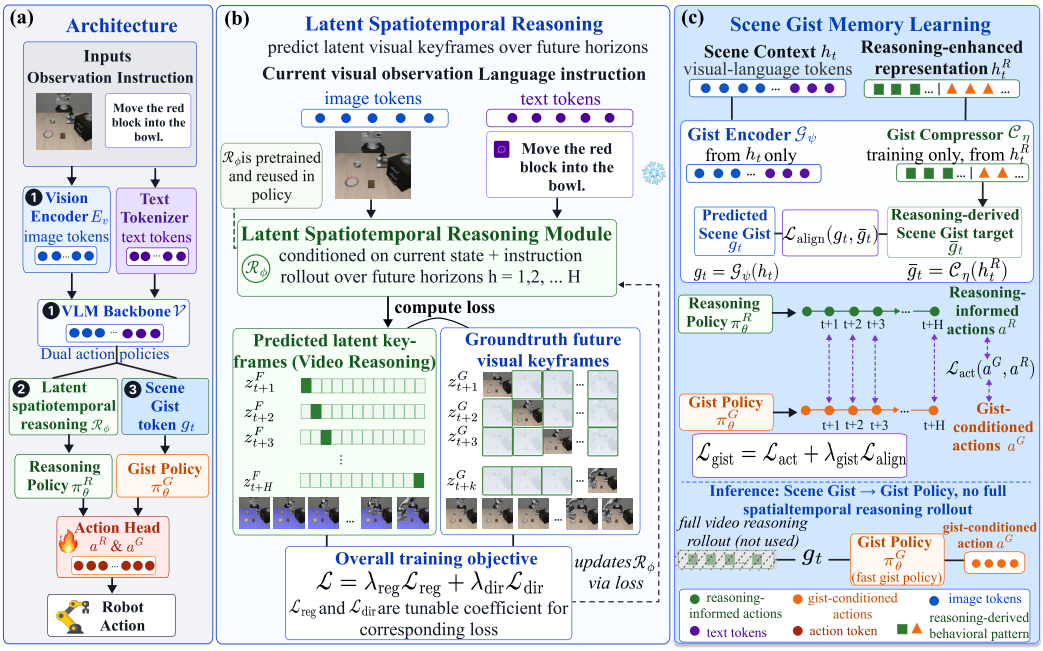

Figure 3: Overview of our framework. **(a)** the overall architecture; **(b)** the latent spatiotemporal reasoning module learns to predict future visual keyframes; **(c)** the gist modules internalize reasoning-derived behavioral information into a compact Scene Gist Token.

❷ Latent Spatiotemporal Reasoning Module and Reasoning Policy. We introduce a latent spatiotemporal reasoning module $\mathcal{R}_\phi$ to anticipate task-relevant scene evolution and guide action trajectory selection. Rather than autoregressively generating pixel-level frames, $\mathcal{R}_\phi$ predicts future states directly in visual representation space:

$$z_{t+k}^{\mathcal{F}} = \mathcal{R}_\phi\left(h_t, L_t, k\right), \qquad k = 1, \ldots, H, \tag{1}$$

where $z_{t+k}^{\mathcal{F}}$ denotes the predicted scene representation at the subsequent k-th future step. Together, these latent states form a compact trajectory of anticipated spatiotemporal dynamics.

Our framework consists of a reasoning policy that predicts the future actions based on the current multimodal representation and the predicted latent future representation:

$$a_{t+k}^R = \pi_\theta^R\left(h_t, L_t, z_{t+k}^F\right). \tag{2}$$

where a_{t+k}^R denotes predicted action at the kth future step. It explicitly exploits anticipated scene evolution to determine future actions and provides reasoning-informed action guidance in training.

❸ Gist Compressor, Gist Encoder, and Gist Policy. We introduce a gist compressor and a gist encoder to summarize reasoning-derived information and predict it directly from the current context, respectively. After the complete latent spatiotemporal reasoning rollout, the implicit action reasoner fuses the current visual–language context with all predicted future representations to obtain a reasoning-enhanced representation h_t^R. The gist compressor $\mathcal{C}_\eta$ maps this representation through an MLP bottleneck into a single Scene Gist vector:

$$\bar{g}_t = \mathcal{C}_\eta(h_t^R). \tag{3}$$

The compressor is trained by matching its induced actions to those of the full reasoning policy, so that $\bar{g}_t$ retains behaviorally relevant information from the complete reasoning pathway and serves as a target for the gist encoder. The gist encoder $\mathcal{G}_\psi$ then learns to predict this target using only the current observation and instruction, represented by h_t:

$$g_t = \mathcal{G}_\psi(h_t). \tag{4}$$

The encoder concatenates the masked means and standard deviations of the current prefix key–value features across layers, normalizes the resulting statistics, and maps them to a gist vector through a

two-layer MLP. The gist policy uses the predicted g_t in place of the implicit action reasoner's output. The existing interaction module incorporates this vector into action generation:

$$a^G_{t:t+H-1} = \pi^G_\theta(h_t, L_t, g_t). \tag{5}$$

The original visual–language prefix and explicit action reasoning module are retained, allowing the policy to combine detailed visual information with the behavioral guidance provided by the gist. At inference, the model predicts the gist directly from the current context, bypassing the latent spatiotemporal reasoning rollout, the implicit action reasoner, and the gist compressor.

3.2 Model Training with Future Reasoning and Scene Gist Memory

Our VLA framework incorporates two core mechanisms, reasoning over possible future predicted trajectories and scene gist tokens. We first train the reasoning policy and then train the Scene Gist modules with the pretrained reasoning policy frozen. The training details are specified below.

3.2.1 Learning Spatiotemporal Future Reasoning For Trajectory Alignment

We first encode the visual observation into a latent visual representation z_t. Instead of explicitly reconstructing future observations at pixel level, we train a spatiotemporal reasoning module $\mathcal{R}_\phi$ to anticipate environment dynamics and potential trajectory directly in visual representation space. Specifically, conditioned on the current visual state and task instruction, the reasoning module predicts

$$z^{\mathcal{F}}_{t+k} = \mathcal{R}_\phi(h_t, L_t, k), \tag{6}$$

and the anticipated future representation can be expressed as $\hat{z}_{t+k} = z_t + z^{\mathcal{F}}_{t+k}$. Here, $z^{\mathcal{F}}_{t+k}$ represents the predicted visual transition from the current state at time t to the kth future step, capturing how the scene is expected to evolve in the latent space. The ground truth is $z^G_{t+k} = z_{t+k} - z_t$. Predicting the visual transition rather than the complete future representation encourages the reasoning module to focus on task-relevant changes in the environment, such as object displacement, robot motion, and interaction events, while reducing the need to reproduce static scene information already contained in the current observation, as shown in Figure 3.

We supervise the predicted future evolution using the ground truth visual transitions extracted from demonstration trajectories. The primary objective minimizes the discrepancy between them:

$$\mathcal{L}_{\text{reg}} = \frac{1}{K}\sum_{k=1}^{K}\left\| z^{\mathcal{F}}_{t+k} - z^G_{t+k} \right\|_2^2. \tag{7}$$

With $\mathcal{L}_{\text{reg}}$ encouraging a magnitude similarity, an additional loss is adopted to encourage directional consistency between the predicted and target transitions:

$$\mathcal{L}_{\text{dir}} = \frac{1}{K}\sum_{k=1}^{K}\left(1 - \cos\left(z^{\mathcal{F}}_{t+k}, z^G_{t+k}\right)\right). \tag{8}$$

3.2.2 Internalizing Spatiotemporal Reasoning into Scene Gist Memory

We train the Scene Gist modules in two stages. First, the gist compressor $\mathcal{C}_\eta$ is trained with the reasoning policy frozen to compress the reasoning-enhanced representation h^R_t into a behaviorally informative target $\bar{g}_t$. Specifically, $\bar{g}_t$ replaces the implicit action reasoner's output, and the compressor is optimized by matching the resulting actions to those of the full reasoning policy. We then freeze the compressor and train the gist encoder to predict $g_t = \mathcal{G}_\psi(h_t)$ directly from the current context. The encoder is supervised by both representation alignment and action consistency:

$$\mathcal{L}_{\text{align}} = \|g_t - \text{sg}(\bar{g}_t)\|_2^2, \qquad \mathcal{L}_{\text{act}} = \frac{1}{H}\sum_{k=0}^{H-1}\left\|a^G_{t+k} - \text{sg}(a^R_{t+k})\right\|_2^2. \tag{9}$$

Here, $\text{sg}(\cdot)$ denotes stop-gradient. $\mathcal{L}_{\text{align}}$ encourages the predicted Scene Gist to match the reasoning-derived target produced by the frozen gist compressor, while $\mathcal{L}_{\text{act}}$ encourages the gist policy to preserve the action behavior of the full reasoning policy.

The final Scene Gist objective is

$$\mathcal{L}_{\text{gist}} = \mathcal{L}_{\text{act}} + \lambda_{\text{gist}} \mathcal{L}_{\text{align}}, \tag{10}$$

where λ_{gist} weights the representation alignment loss relative to the action consistency loss.

Efficient Inference with Gist Token. At inference time, the gist encoder directly predicts a single Scene Gist g_t from the current context and uses it to condition the existing action-generation pathway for the entire action chunk. The latent reasoning rollout, implicit action reasoner, and gist compressor are therefore bypassed, substantially reducing inference overhead.

4 Experiments

4.1 Experimental Setup

Dataset. We use LIBERO Liu et al. (2023) as the primary training benchmark and evaluate success rates on the Spatial, Object, Goal, and Long suites. We further evaluate robustness on LIBERO-Plus Fei et al. (2026) under zero-shot transfer and supervised fine-tuning, where the former directly transfers the LIBERO-trained policy to perturbed environments and the latter further trains it on LIBERO-Plus. Finally, we evaluate on VLABench Zhang et al. (2025a) following Zhong et al. (2026), where the Non-Latent Spatiotemporal Reasoning baseline and our latent spatiotemporal reasoning module are fine-tuned on identical training data and evaluated on five public tracks: in-distribution, cross-category, commonsense, semantic-instruction, and unseen-texture.

Baselines. We compare our method with representative VLA and world-model approaches, including ACoT-VLA Zhong et al. (2026), Octo Mees et al. (2024), OpenVLA Kim et al. (2024), π_0 Black et al. (2024), $\pi_{0.5}$ Physical Intelligence et al. (2025), CoT-VLA Zhao et al. (2025a), WorldVLA Cen et al. (2025), DreamVLA Zhang et al. (2025b), Fast-WAM Yuan et al. (2026), and ImageWAM Zhang et al. (2026). The complete comparison sets for each benchmark are reported in Tables 1–3, with reproduced results indicated in the corresponding table captions.

Implementation Details. We use 8 NVIDIA H100 GPUs to train the reasoning module and use 4 NVIDIA A6000 GPUs to train the gist memeory module, while we only use a single NVIDIA A6000 GPU for end-to-end inference latency evaluation. We build on the PaliGemma-based backbone Zhong et al. (2026), using SigLIP So400m/14 as the vision encoder, Gemma-2B as the multimodal prefix model, and Gemma-300M modules as action experts. Latent spatiotemporal reasoning module predicts $K = 10$ future SigLIP latent states, supervised by latent MSE and cosine direction losses, while the auxiliary pixel reconstruction head is used only during training. For LIBERO, we train for 100K steps with a batch size of 16 and a peak learning rate of 5×10^{-6} (cosine decay to 1×10^{-7}). For VLABench, we follow the configuration of Zhong et al. (2026) with 61K training steps and a batch size of 128 to ensure a controlled comparison.

4.2 Simulation Experiments

Main Results. Tables 1–3 present the simulation results on LIBERO, LIBERO-Plus, and VLABench. On LIBERO, both IG-VLA (Reasoning Policy) and IG-VLA$^{\triangle}$ (Gist Policy) achieve the highest average success rate of 98.9%, outperforming the strongest baseline ACoT-VLA of 98.5%. In particular, IG-VLA$^{\triangle}$ achieves 100.0% and 97.8% success rates on the Object and Long suites, respectively, demonstrating that Scene Gist preserves strong manipulation performance without explicit future reasoning rollout. On LIBERO-Plus, our supervised fine-tuning achieves the best average success rate of 88.5%, exceeding best baseline ACoT-VLA by 0.5 percentage points, with a notable improvement on language variation from 79.7% to 85.6%. Consistent improvements are also observed on VLABench, where our method achieves the highest average Intention Score (IS) of 66.68, improving over ACoT-VLA by 3.18 points. The improvement is particularly pronounced under semantic-instruction variation, where IS and PS increase from 56.8/39.6 to 65.2/48.49, respectively. Although the average PS on VLABench is slightly lower than ACoT-VLA (46.11 vs. 47.40), the overall results across the three benchmarks demonstrate that latent spatiotemporal reasoning improves intention understanding and action prediction, while Scene Gist effectively internalizes these reasoning benefits for efficient manipulation without explicit future rollout.

Table 1: Comparison on the LIBERO benchmark. * represents that the LLM backbone is frozen during training. Our proposed approach is trained on the LIBERO dataset. All metrics are average success rates (%). The best results are highlighted in **bold**. △ denotes IG-VLA with Latent Spatiotemporal Reasoning replaced by Scene Gist Memory.

Methods	Spatial		Object		Goal		Long		Avg.	
	SR ↑	Rank ↓	SR ↑	Rank ↓	SR ↑	Rank ↓	SR ↑	Rank ↓	SR ↑	Rank ↓
Diffusion Policy	78.3	29	92.5	21	68.3	30	50.5	30	72.4	30
Octo	78.9	28	85.7	29	84.6	24	51.1	29	75.1	28
CoT-VLA	87.5	23	91.6	23	87.6	22	69.0	22	81.1	23
WorldVLA (8) (256×256)	85.6	25	89.0	26	82.6	26	59.0	25	79.1	24
WorldVLA (8) (512×512)	87.6	22	96.2	18	83.4	25	60.0	24	81.8	22
DreamVLA	97.5	11	94.0	19	89.5	18	89.5	15	92.6	17
UniVLA	95.4	17	98.8	7	93.6	15	94.0	9	95.5	13
F1	98.2	6	97.8	13	95.4	14	91.3	13	95.7	12
GE-Act	98.2	6	97.6	15	95.8	12	94.4	8	96.5	10
Fast-WAM	98.2	6	100.0	1	97.0	10	95.2	5	97.6	5
ImageWAM	97.2	12	99.2	4	98.8	4	98.4	1	98.4	4
TraceVLA	84.6	27	85.2	30	75.1	29	54.1	27	74.8	29
OpenVLA	84.7	26	88.4	27	79.2	27	53.7	28	76.5	27
UniAct	77.0	30	87.0	28	77.0	28	70.0	21	77.8	26
SpatialVLA	88.2	21	89.9	25	88.6	20	55.5	26	78.1	25
ThinkAct	88.3	20	91.4	24	87.1	23	70.9	20	84.4	21
π_0-FAST	96.4	15	96.8	17	88.6	20	60.2	23	85.5	20
FPC-VLA	87.0	24	92.0	22	89.2	19	82.2	18	86.9	19
SmoVLA	93.0	19	94.0	19	91.0	17	77.0	19	88.8	18
GR00T-N1	94.4	18	97.6	15	93.0	16	90.6	14	93.9	16
π_0	96.8	14	98.8	7	95.8	12	85.2	17	94.1	15
GO-1	96.2	16	97.8	13	96.0	11	89.2	16	94.8	14
DD-VLA	97.2	12	98.6	9	97.4	8	92.0	12	96.3	11
MemoryVLA	98.4	5	98.4	10	97.4	8	93.4	10	96.7	9
$\pi_{0.5}$	98.8	3	98.2	12	98.0	6	92.4	11	96.9	8
OpenVLA-OFT	97.6	10	98.4	10	97.9	7	94.5	7	97.1	7
VLA-Adapter	97.8	9	99.2	4	99.2	2	95.0	6	97.3	6
ACoT-VLA*	99.4	1	99.6	3	98.8	4	96.0	4	98.5	2
ACoT-VLA*	98.6	4	99.0	6	99.4	1	97.0	3	98.5	2
IG-VLA (Ours)	**99.4**	1	**99.8**	2	99.2	2	**97.2**	2	**98.9**	1
IG-VLA (Ours)$^{\triangle}$	**99.4**	1	**100.0**	1	98.2	5	**97.8**	2	**98.9**	1

Table 2: Supervised fine-tuning results on LIBERO-Plus. ◇ denotes a pipeline with a frozen LLM backbone, and △ denotes IG-VLA with Latent Spatiotemporal Reasoning replaced by Scene Gist Memory. Best results are in **bold**.

Methods	Camera	Robot	Language	Light	Background	Noise	Layout	Avg.
$\pi_0^{\diamond}$	79.6	21.1	72.5	84.7	86.2	68.3	69.4	67.4
$\pi_{0.5}^{\diamond}$	70.3	41.7	81.1	97.3	94.6	71.8	84.9	75.7
ACoT-VLA$^{\diamond}$	91.2	62.5	80.3	95.1	91.5	88.3	84.9	84.1
ACoT-VLA	96.6	70.4	79.7	95.1	97.1	95.9	85.0	88.0
IG-VLA (Ours)	95.4	68.9	**85.6**	**97.5**	97.2	94.9	**85.5**	**88.5**
IG-VLA (Ours)$^{\triangle}$	95.3	69.0	85.1	97.4	**97.5**	94.7	84.6	88.3

Table 3: Comparison on VLABench. IS and PS denote Intention Score and Progress Score. ◇ denotes pipeline with a frozen LLM backbone, and △ denotes IG-VLA with Latent Spatiotemporal Reasoning replaced by Scene Gist Memory. Best results are in **bold**.

Methods	Guidance	In-dist.		Category		Commonsense		Instruction		Texture		Avg.	
		IS ↑	PS ↑	IS ↑	PS ↑	IS ↑	PS ↑	IS ↑	PS ↑	IS ↑	PS ↑	IS ↑	PS ↑
$\pi_0^{\diamond}$	Linguistics	67.8	62.7	44.0	33.6	54.9	**43.0**	58.0	38.7	50.6	42.5	55.0	44.1
$\pi_{0.5}^{\diamond}$	Linguistics	75.0	60.8	49.6	35.3	57.5	41.6	57.1	30.3	62.0	47.4	60.2	43.1
ACoT-VLA	**Action**	79.8	**66.1**	54.1	**38.9**	52.3	37.8	56.8	39.6	74.6	**54.6**	63.5	**47.4**
IG-VLA (Ours)	**Action**	**80.80**	61.64	54.20	32.21	**58.20**	38.97	**65.20**	**48.49**	**74.99**	49.23	**66.68**	46.11
IG-VLA (Ours)$^{\triangle}$	**Action**	80.48	61.23	**54.35**	31.75	57.82	39.18	64.76	48.22	74.71	48.88	66.26	45.68

End-to-end Inference Speedup. Table 4 compares the end-to-end inference latency of our method with existing baselines on LIBERO using a single NVIDIA A6000 GPU. We report latency per action chunk to characterize the computational cost of generating each action sequence. With Scene

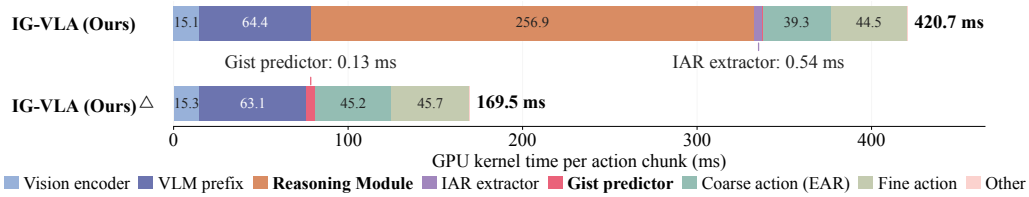

Figure 4: GPU time breakdown(batch size 1, BF16), averaged over three profiled forward passes.

Gist Memory, IG-VLA$^{\triangle}$ requires only 169.5 ms per action chunk, achieving $1.78\times$ and $1.55\times$ speedups over FastWAM (1-step) (302 ms) and ImageWAM (263 ms), respectively. Alongside this inference efficiency, the gist policy achieves a 98.9% average success rate on LIBERO, surpassing ImageWAM (98.4%) and the strongest baseline, ACoT-VLA (98.5%). These results indicate that the lower inference cost is accompanied by strong manipulation performance. By internalizing reasoning-derived scene–behavior associations into a compact Scene Gist Token, the gist policy guides action prediction without repeatedly performing explicit future rollouts. This allows IG-VLA$^{\triangle}$ to retain the benefits of future reasoning while reducing its inference overhead, supporting efficient VLA deployment. Baseline results are taken from the ImageWAM paper (Zhang et al., 2026).

Table 4: Inference latency comparison with baselines using a single NVIDIA A6000 GPU.

Method	Test-time Future Representation	Latency ↓
FastWAM-IDM	Generated video	1081 ms
FastWAM (1-step)	Cached video features	302 ms
ImageWAM	Cached image features	263 ms
IG-VLA (Ours)	Latent future rollout	420.7 ms
IG-VLA (Ours)$^{\triangle}$	Scene gist token	**169.5 ms**

Table 5: Ablation of different numbers of predicted latent visual keyframes K on LIBERO.

Suite	$K=3$	$K=5$	$K=10$	$K=15$
LIBERO-Spatial	98.4	98.6	**99.4**	98.0
LIBERO-Object	**99.8**	99.6	**99.8**	99.2
LIBERO-Goal	97.4	98.8	**99.2**	98.4
LIBERO-10	96.8	96.4	**97.2**	95.6
Average	98.10	98.35	**98.90**	97.80

Latency Breakdown Analysis. Figure 4 further analyzes the computational cost of our reasoning and gist policies. Explicit spatiotemporal reasoning incurs 256.9 ms of GPU time, whereas the Scene Gist predictor introduces only 0.13 ms of additional overhead. Consequently, replacing explicit future reasoning with Scene Gist reduces the overall GPU time from 420.7 ms to 169.5 ms. Importantly, this substantial efficiency gain does not compromise task performance: our gist policy achieves the highest average success rate of 98.9% on LIBERO, outperforming the strongest baseline ACoT-VLA (98.5%) by 0.4 percentage points.

4.3 Ablation Study

Effects of Different Number of Latent Keyframes. Table 5 studies the effect of different number of predicted latent spatiotemporal keyframes K on LIBERO. Increasing K from 3 to 10 consistently improves the average success rate from 98.10% to 98.90%, suggesting that a sufficient imagination horizon provides richer future dynamics for action reasoning. However, further increasing K to 15 reduces the success rate to 97.80%. This degradation can be attributed to longer-horizon future prediction, which can accumulate prediction errors and introduce redundant or unreliable future information Huang et al. (2026c); Jin et al. (2026), thereby weakening its guidance for downstream action reasoning. We therefore use $K = 10$ as the default setting in main experiments.

Effect of λ_{dir}. We study the relative weighting between the two training objectives by fixing $\lambda_{\text{reg}} = 1$ and varying λ_{dir}, thereby controlling the contribution of directional consistency relative to latent transition regression. As shown in Figure 5, $\lambda_{\text{dir}} = 0.1$ (main results default setting) achieves the best success rate of 99.4% on LIBERO-Spatial. Increasing λ_{dir} to 0.3 slightly reduces the success rate to 98.6%, while larger weights of 0.5 and 0.8 lead to substantial degradation to 37.0% and 25.7%, respectively. This suggests that the directional objective is most effective as a lightweight auxiliary constraint. A moderate weight encourages alignment with the future visual transition, whereas an overly large weight overemphasizes directional consistency relative to transition-magnitude regression and destabilizes the learning objective. We therefore set $\lambda_{\text{dir}} = 0.1$ and $\lambda_{\text{reg}} = 1$ in our main experiments.

Effects of Different Number of MLP Layers for Training Gist Policy. We ablate the number of MLP layers in the Gist Policy to study the trade-off between model capacity and training efficiency.

As shown in Figure 6, increasing the depth from one to two layers improves the average LIBERO success rate from 97.8% to 98.9%, indicating that a single layer provides insufficient capacity to effectively capture the reasoning-derived Scene Gist. Further increasing the depth to three layers provides no additional benefit, slightly decreasing the success rate to 98.8% while increasing the training cost from 3.45 to 5.79 GPU hours. We therefore adopt a two-layer MLP for the Gist Policy, which provides sufficient capacity while avoiding unnecessary training overhead.

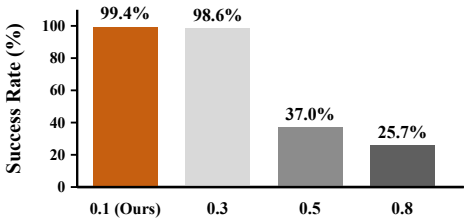

Figure 5: Ablation of the directional consistency loss weight λ_{dir} on LIBERO-Spatial.

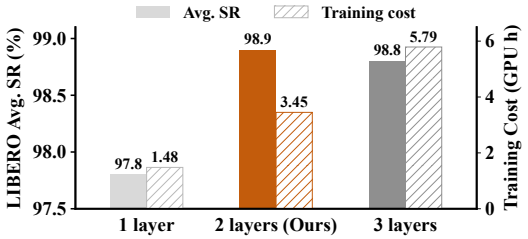

Figure 6: Effects of different layers of MLP for training gist policy.

4.4 Real-World Deployment

Real-World Evaluation. We further deploy the fine-tuned policy on a dual-arm UR3 robotic platform equipped with parallel-jaw grippers and three RGB cameras, as shown in Figure 7(a). We consider a pick-and-place task in which the robot grasps a screwdriver and places it at a target location, and fine-tune all policies on 77 teleoperated demonstrations. Under identical deployment and fine-tuning settings, we compare our method with $\pi_{0.5}$ and ACoT-VLA, conducting 16 trials for each policy with randomized initial screwdriver placements. As shown in Figure 7(b), our method achieves the highest success rate of 62.5%, compared with 56.2% for $\pi_{0.5}$ and 50.0% for ACoT-VLA. These results further demonstrate that incorporating future spatiotemporal reasoning improves manipulation performance beyond the underlying VLA backbone and prior action-reasoning approaches. Detailed deployment settings and action-space conversion are provided in Appendix A.2.

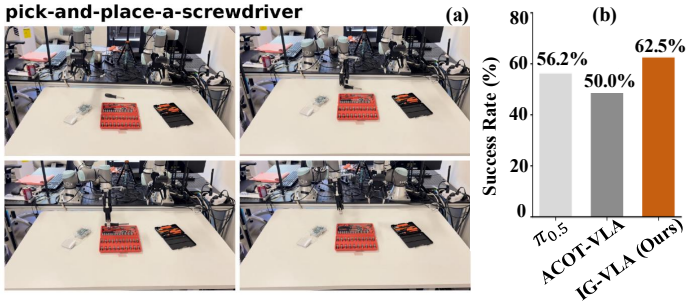

Figure 7: Real-world evaluation on the dual-arm UR3. (a) Pick-and-place-a-screwdriver task execution. (b) Success rate comparison with baselines over 16 trials per method.

5 Conclusion

In this work, we present IG-VLA to address two key limitations of existing VLA reasoning: the lack of explicit anticipation of future scene dynamics and the high inference cost of repeated reasoning rollouts. By reasoning over future dynamics in latent visual space, IG-VLA provides task-relevant spatiotemporal guidance for action generation, while Scene Gist Memory further converts the acquired reasoning knowledge into compact behavioral guidance for efficient inference. Experiments across LIBERO, LIBERO-Plus, and VLABench consistently demonstrate improved manipulation performance and generalization, with particularly strong gains under challenging language and semantic variations. Meanwhile, the gist policy retains competitive reasoning performance while substantially reducing inference cost, showing that explicit future reasoning need not be repeatedly executed at deployment. Real-world robotic experiments further validate the effectiveness of the proposed framework beyond simulation. Overall, our results suggest that anticipating future scene dynamics and subsequently internalizing such reasoning provides an effective direction toward more capable and efficient VLA policies.

AI USE STATEMENT

In this work, large language models (LLMs) were used solely as a general-purpose writing assistant. Their role was limited to correcting grammar, fixing typographical errors, and polishing the language for clarity and readability.

REFERENCES

Cristina M Atance and Daniela K O'Neill. Episodic future thinking. *Trends in cognitive sciences*, 5 (12):533–539, 2001.

Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, et al. backslashpi0 : A vision-language-action flow model for general robot control. *arXiv preprint arXiv:2410.24164*, 2024.

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Jasmine Hsu, et al. Rt-1: Robotics transformer for real-world control at scale. *arXiv preprint arXiv:2212.06817*, 2022.

Jun Cen, Chaohui Yu, Hangjie Yuan, Yuming Jiang, Siteng Huang, Jiayan Guo, Xin Li, Yibing Song, Hao Luo, Fan Wang, Deli Zhao, and Hao Chen. Worldvla: Towards autoregressive action world model. *arXiv preprint arXiv:2506.21539*, 2025.

Jaden Clark, Suvir Mirchandani, Dorsa Sadigh, and Suneel Belkhale. Action-free reasoning for policy generalization. *arXiv preprint arXiv:2502.03729*, 2025.

Federica Conti and Muireann Irish. Harnessing visual imagery and oculomotor behaviour to understand prospection. *Trends in Cognitive Sciences*, 25(4):272–283, 2021.

Danny Driess, Jost Springenberg, Brian Ichter, Lili Yu, Adrian Li-Bell, Karl Pertsch, Allen Ren, Homer Walke, Quan Vuong, Lucy Xiaoyang Shi, et al. Knowledge insulating vision-language-action models: Train fast, run fast, generalize better. *Advances in Neural Information Processing Systems*, 38:102867–102888, 2026.

Senyu Fei, Siyin Wang, Junhao Shi, Zihao Dai, Jikun Cai, Pengfang Qian, Li Ji, Xinzhe He, Shiduo Zhang, Zhaoye Fei, Jinlan Fu, Jingjing Gong, and Xipeng Qiu. LIBERO-Plus: A progressive robustness benchmark for visual-language-action models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 38574–38583, 2026.

Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, Junfei Wu, Xiaoying Zhang, Benyou Wang, and Xiangyu Yue. Video-r1: Reinforcing video reasoning in mllms. *Advances in Neural Information Processing Systems*, 38:99114–99137, 2026.

Weifan Guan, Qinghao Hu, Aosheng Li, and Jian Cheng. Efficient vision-language-action models for embodied manipulation: A systematic survey. *arXiv preprint arXiv:2510.17111*, 2025.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025a.

Wenkai Guo, Guanxing Lu, Haoyuan Deng, Zhenyu Wu, Yansong Tang, and Ziwei Wang. Vla-reasoner: Empowering vision-language-action models with reasoning via online monte carlo tree search. *arXiv preprint arXiv:2509.22643*, 2025b.

Chi-Pin Huang, Yunze Man, Zhiding Yu, Min-Hung Chen, Jan Kautz, Yu-Chiang Frank Wang, and Fu-En Yang. Fast-thinkact: Efficient vision-language-action reasoning via verbalizable latent planning. *arXiv preprint arXiv:2601.09708*, 2026a.

Chi-Pin Huang, Yueh-Hua Wu, Min-Hung Chen, Frank Wang, and Fu-En Yang. Thinkact: Vision-language-action reasoning via reinforced visual latent planning. *Advances in Neural Information Processing Systems*, 38:82782–82802, 2026b.

Junchao Huang, Ziyang Ye, Xinting Hu, Tianyu He, Guiyu Zhang, Shaoshuai Shi, Jiang Bian, and Li Jiang. Live: Long-horizon interactive video world modeling. *arXiv preprint arXiv:2602.03747*, 2026c.

Minghao Jin, Mozheng Liao, Mingfei Han, Zhihui Li, and Xiaojun Chang. Beyond dense futures: World models as structured planners for robotic manipulation. *arXiv preprint arXiv:2603.12553*, 2026.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Pannag Sanketi, Quan Vuong, Thomas Kollar, Benjamin Burchfiel, Russ Tedrake, Dorsa Sadigh, Chelsea Finn, Percy Liang, and Sergey Levine. Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*, 2024.

Jason Lee, Jiafei Duan, Haoquan Fang, Yuquan Deng, Shuo Liu, Boyang Li, Bohan Fang, Jieyu Zhang, Yi Ru Wang, Sangho Lee, et al. Molmoact: Action reasoning models that can reason in space. *arXiv preprint arXiv:2508.07917*, 2025.

Hengtao Li, Pengxiang Ding, Runze Suo, Yihao Wang, Zirui Ge, Dongyuan Zang, Kexian Yu, Mingyang Sun, Hongyin Zhang, Donglin Wang, et al. Vla-rft: Vision-language-action reinforcement fine-tuning with verified rewards in world simulators. *arXiv preprint arXiv:2510.00406*, 2025a.

Xinqing Li, Xin He, Le Zhang, Min Wu, Xiaoli Li, and Yun Liu. A comprehensive survey on world models for embodied ai. *arXiv preprint arXiv:2510.16732*, 2025b.

Yi Li, Yuquan Deng, Jesse Zhang, Joel Jang, Marius Memmel, Caelan Garrett, Fabio Ramos, Dieter Fox, Anqi Li, Abhishek Gupta, et al. Hamster: Hierarchical action models for open-world robot manipulation. In *International Conference on Learning Representations*, volume 2025, pp. 24040–24068, 2025c.

Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qiang Liu, Yuke Zhu, and Peter Stone. LIBERO: Benchmarking knowledge transfer for lifelong robot learning. In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track*, 2023.

Yang Liu, Weixing Chen, Yongjie Bai, Xiaodan Liang, Guanbin Li, Wen Gao, and Liang Lin. Aligning cyber space with physical world: A comprehensive survey on embodied ai. *IEEE/ASME Transactions on Mechatronics*, 2025a.

Ziyu Liu, Zeyi Sun, Yuhang Zang, Xiaoyi Dong, Yuhang Cao, Haodong Duan, Dahua Lin, and Jiaqi Wang. Visual-rft: Visual reinforcement fine-tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2034–2044, 2025b.

Yueen Ma, Zixing Song, Yuzheng Zhuang, Jianye Hao, and Irwin King. A survey on vision–language–action models for embodied ai. *IEEE Transactions on Neural Networks and Learning Systems*, 2026.

Oier Mees, Dibya Ghosh, Karl Pertsch, Kevin Black, Homer Rich Walke, Sudeep Dasari, Joey Hejna, Tobias Kreiman, Charles Xu, Jianlan Luo, et al. Octo: An open-source generalist robot policy. In *First Workshop on Vision-Language Models for Navigation and Manipulation at ICRA 2024*, 2024.

J Bjorck Nvidia, Fernando Castaneda, N Cherniadev, X Da, R Ding, L Fan, Y Fang, D Fox, F Hu, S Huang, et al. Gr00t n1: An open foundation model for generalist humanoid robots. *arXiv preprint arXiv:2503.14734*, 2, 2025.

Karl Pertsch, Kyle Stachowicz, Brian Ichter, Danny Driess, Suraj Nair, Quan Vuong, Oier Mees, Chelsea Finn, and Sergey Levine. Fast: Efficient action tokenization for vision-language-action models. *arXiv preprint arXiv:2501.09747*, 2025.

Physical Intelligence, Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Manuel Y. Galliker, Dibya Ghosh, Lachy Groom, Karol Hausman, Brian Ichter, Szymon Jakubczak, Tim Jones, Liyiming Ke, Devin LeBlanc, Sergey Levine, Adrian Li-Bell, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Allen Z. Ren, Lucy Xiaoyang Shi, Laura Smith, Jost Tobias Springenberg, Kyle Stachowicz, James Tanner, Quan Vuong, Homer Walke, Anna Walling, Haohuan Wang, Lili Yu, and Ury Zhilinsky. $\pi_{0.5}$: A vision-language-action model with open-world generalization, 2025.

Delin Qu, Haoming Song, Qizhi Chen, Zhaoqing Chen, Xianqiang Gao, Xinyi Ye, Qi Lv, Modi Shi, Guanghui Ren, Cheng Ruan, et al. Eo-1: Interleaved vision-text-action pretraining for general robot control. arxiv preprint, 2025. *URL https://arxiv. org/abs/2508.21112*, 2.

Ranjan Sapkota, Yang Cao, Konstantinos I Roumeliotis, and Manoj Karkee. Vision-language-action (vla) models: Concepts, progress, applications and challenges. *arXiv preprint arXiv:2505.04769*, 2025.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Huajie Tan, Yuheng Ji, Xiaoshuai Hao, Minglan Lin, Pengwei Wang, Zhongyuan Wang, and Shanghang Zhang. Reason-rft: Reinforcement fine-tuning for visual reasoning. *arXiv e-prints*, pp. arXiv–2503, 2025.

Chaoyang Wang, Zeyu Zhang, Meng Meng, Xu Zhou, and Haiyun Jiang. Vision-ekipl: External knowledge-infused policy learning for visual reasoning. *arXiv preprint arXiv:2506.06856*, 2025.

Yihao Wang, Pengxiang Ding, Lingxiao Li, Can Cui, Zirui Ge, Xinyang Tong, Wenxuan Song, Han Zhao, Wei Zhao, Pengxu Hou, et al. Vla-adapter: An effective paradigm for tiny-scale vision-language-action model. In *Proceedings of the AAAI conference on artificial intelligence*, volume 40, pp. 18638–18646, 2026.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Wendy Wood and Dennis Rünger. Psychology of habit. *Annual review of psychology*, 67(1):289–314, 2016.

Gabriele Wulf, Charles Shea, and Rebecca Lewthwaite. Motor skill learning and performance: a review of influential factors. *Medical education*, 44(1):75–84, 2010.

Jianwei Yang, Reuben Tan, Qianhui Wu, Ruijie Zheng, Baolin Peng, Yongyuan Liang, Yu Gu, Mu Cai, Seonghyeon Ye, Joel Jang, et al. Magma: A foundation model for multimodal ai agents. In *Proceedings of the computer vision and pattern recognition conference*, pp. 14203–14214, 2025.

Angen Ye, Zeyu Zhang, Boyuan Wang, Xiaofeng Wang, Dapeng Zhang, and Zheng Zhu. Vla-r1: Enhancing reasoning in vision-language-action models. *arXiv preprint arXiv:2510.01623*, 2025.

Edward Yeo, Yuxuan Tong, Morry Niu, Graham Neubig, and Xiang Yue. Demystifying long chain-of-thought reasoning in llms. *arXiv preprint arXiv:2502.03373*, 2025.

Cheng Yin, Yankai Lin, Wang Xu, Sikyuen Tam, Xiangrui Zeng, Zhiyuan Liu, and Zhouping Yin. Deepthinkvla: Enhancing reasoning capability of vision-language-action models. *arXiv preprint arXiv:2511.15669*, 2025.

Tianyuan Yuan, Zibin Dong, Yicheng Liu, and Hang Zhao. Fast-wam: Do world action models need test-time future imagination? *arXiv preprint arXiv:2603.16666*, 2026. URL https://arxiv.org/abs/2603.16666.

Michał Zawalski, William Chen, Karl Pertsch, Oier Mees, Chelsea Finn, and Sergey Levine. Robotic control via embodied chain-of-thought reasoning. *arXiv preprint arXiv:2407.08693*, 2024.

Shiduo Zhang, Zhe Xu, Peiju Liu, Xiaopeng Yu, Yuan Li, Qinghui Gao, Zhaoye Fei, Zhangyue Yin, Zuxuan Wu, Yu-Gang Jiang, et al. Vlabench: A large-scale benchmark for language-conditioned robotics manipulation with long-horizon reasoning tasks. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11142–11152. IEEE, 2025a.

Wenyao Zhang, Hongsi Liu, Zekun Qi, Yunnan Wang, Xinqiang Yu, Jiazhao Zhang, Runpei Dong, Jiawei He, Fan Lu, He Wang, Zhizheng Zhang, Li Yi, Wenjun Zeng, and Xin Jin. Dreamvla: A vision-language-action model dreamed with comprehensive world knowledge. *arXiv preprint arXiv:2507.04447*, 2025b.

Yuyang Zhang, Wenyao Zhang, Zekun Qi, He Zhang, Haitao Lin, Jingbo Zhang, Yao Mu, Xiaokang Yang, Wenjun Zeng, and Xin Jin. Imagewam: Do world action models really need video generation, or just image editing? *arXiv preprint arXiv:2606.19531*, 2026. URL https://arxiv.org/abs/2606.19531.

Qingqing Zhao, Yao Lu, Moo Jin Kim, Zipeng Fu, Zhuoyang Zhang, Yecheng Wu, Zhaoshuo Li, Qianli Ma, Song Han, Chelsea Finn, Ankur Handa, Tsung-Yi Lin, Gordon Wetzstein, Ming-Yu Liu, and Donglai Xiang. Cot-vla: Visual chain-of-thought reasoning for vision-language-action models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1702–1713, 2025a.

Qingqing Zhao, Yao Lu, Moo Jin Kim, Zipeng Fu, Zhuoyang Zhang, Yecheng Wu, Zhaoshuo Li, Qianli Ma, Song Han, Chelsea Finn, et al. Cot-vla: Visual chain-of-thought reasoning for vision-language-action models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 1702–1713, 2025b.

Ruijie Zheng, Yongyuan Liang, Shuaiyi Huang, Jianfeng Gao, Hal Daumé III, Andrey Kolobov, Furong Huang, and Jianwei Yang. Tracevla: Visual trace prompting enhances spatial-temporal awareness for generalist robotic policies. In *International Conference on Learning Representations*, volume 2025, pp. 54277–54296, 2025.

Linqing Zhong, Yi Liu, Yifei Wei, Ziyu Xiong, Si Liu, and Guanghui Ren. Acot-vla: Action chain-of-thought for vision-language-action models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8152–8162, 2026.

A Appendix

A.1 Overall Detailed Pipeline

Algorithm 1 Learning and Inference with Latent Spatiotemporal Reasoning and Scene Gist

Require: Demonstration dataset $\mathcal{D}$, vision encoder E_v, vision-language backbone $\mathcal{V}$, spatiotemporal reasoning module $\mathcal{R}_\phi$, reasoning policy π_θ^R, gist compressor $\mathcal{C}_\eta$, gist encoder $\mathcal{G}_\psi$, gist policy π_θ^G, reasoning horizon H

Training
for each training sample $(o_t, L_t, \{o_{t+k}, a_{t+k}\}_{k=1}^H) \sim \mathcal{D}$ **do**
  $z_t \leftarrow E_v(o_t), \quad h_t \leftarrow \mathcal{V}(z_t, L_t)$ ▷ Encode current observation
  for $k = 1, \dots, H$ **do**
    $z_{t+k} \leftarrow E_v(o_{t+k})$ ▷ Encode ground-truth future observation
    $z_{t+k}^G \leftarrow z_{t+k} - z_t$ ▷ Construct ground-truth visual transition
    $z_{t+k}^F \leftarrow \mathcal{R}_\phi(h_t, L_t, k)$ ▷ Predict future visual transition
    $\hat{z}_{t+k} \leftarrow z_t + z_{t+k}^F$ ▷ Recover anticipated future representation
    $a_{t+k}^R \leftarrow \pi_\theta^R(h_t, L_t, \hat{z}_{t+k})$ ▷ Predict reasoning-informed action
  end for
  $\mathcal{L}_{\text{reg}} = \frac{1}{H} \sum_{k=1}^{H} \left\| z_{t+k}^F - z_{t+k}^G \right\|_2^2$ ▷ Visual transition regression
  $\mathcal{L}_{\text{dir}} = \frac{1}{H} \sum_{k=1}^{H} \left(1 - \cos(z_{t+k}^F, z_{t+k}^G)\right)$ ▷ Directional consistency
  $h_t^R \leftarrow \text{IAR}\left(h_t, \{\hat{z}_{t+k}\}_{k=1}^H\right)$ ▷ Obtain reasoning-enhanced representation
  $\bar{g}_t \leftarrow \mathcal{C}_\eta(h_t^R)$ ▷ Construct reasoning-derived Scene Gist target
  $\bar{a}_{t:t+H-1}^G \leftarrow \pi_\theta^G(h_t, L_t, \bar{g}_t)$ ▷ Predict actions with Scene Gist target
  $\mathcal{L}_{\text{comp}} = \frac{1}{H} \sum_{k=0}^{H-1} \left\| \bar{a}_{t+k}^G - \text{sg}(a_{t+k}^R) \right\|_2^2$ ▷ Train gist compressor
  $g_t \leftarrow \mathcal{G}_\psi(h_t)$ ▷ Predict Scene Gist from current context
  $a_{t:t+H-1}^G \leftarrow \pi_\theta^G(h_t, L_t, g_t)$ ▷ Predict gist-conditioned action chunk
  $\mathcal{L}_{\text{align}} = \|g_t - \text{sg}(\bar{g}_t)\|_2^2$ ▷ Align predicted and target Scene Gists
  $\mathcal{L}_{\text{act}} = \frac{1}{H} \sum_{k=0}^{H-1} \left\| a_{t+k}^G - \text{sg}(a_{t+k}^R) \right\|_2^2$ ▷ Align reasoning-policy actions
  $\mathcal{L}_{\text{gist}} = \mathcal{L}_{\text{act}} + \lambda_{\text{gist}} \mathcal{L}_{\text{align}}$ ▷ Train Scene Gist encoder
end for
Inference
$z_t \leftarrow E_v(o_t), \quad h_t \leftarrow \mathcal{V}(z_t, L_t)$ ▷ Encode current observation and instruction
$g_t \leftarrow \mathcal{G}_\psi(h_t)$ ▷ Retrieve Scene Gist
$a_{t:t+H-1}^G \leftarrow \pi_\theta^G(h_t, L_t, g_t)$ ▷ Predict action chunk without future rollout
return $a_{t:t+H-1}^G$

Algorithm 1 summarizes the overall training and inference pipeline of IG-VLA. During training, we first encode the current and future observations into the latent visual space and construct their latent transitions as supervision. The Latent Spatiotemporal Reasoning module learns to predict these future transitions from the current multimodal context, enabling the reasoning policy to anticipate task-relevant scene evolution and use the resulting future information for action prediction.

We then internalize the acquired future reasoning into Scene Gist Memory. Specifically, the reasoning-enhanced representation is compressed into a compact Scene Gist target, whose associated actions are aligned with those produced by the reasoning policy. A lightweight gist encoder subsequently learns to predict this reasoning-derived Scene Gist directly from the current multimodal context, with both representation- and action-level supervision. In this way, the Scene Gist captures the behavioral knowledge distilled from explicit future reasoning without requiring the future rollout itself.

At inference time, only the current observation and instruction are required. The gist encoder directly retrieves the corresponding Scene Gist from the current context, which conditions the gist policy to generate the action chunk. Therefore, the expensive Latent Spatiotemporal Reasoning module and future-state rollout are completely bypassed during deployment, while their reasoning benefits are retained through the learned Scene Gist.

A.2 More Details about Real-World Deployment

Setup. To evaluate our method beyond simulation and on an embodiment different from the training platform, we deploy the fine-tuned policy on a dual-arm UR3 setup (Figure **??**). The hardware consists of two 6-DoF UR3 arms with parallel-jaw grippers, yielding a 14-dimensional state/action space, and observations from three RGB cameras (one top-down scene camera and two wrist cameras), each resized to 224×224 before the SigLIP encoder. We consider a pick-and-place-a-screwdriver task: starting from the robot home pose, the right arm grasps a screwdriver and places it at a target location (a single-arm task; the left gripper is disabled during deployment). All policies are fine-tuned from `pi05_base` on 77 teleoperated demonstrations (15 Hz).

Deployment. Since the UR3 is joint-controlled while the policy predicts actions in a delta end-effector (delta-EEF) space, we integrate each predicted delta-EEF chunk onto the current end-effector pose (obtained by forward kinematics from the measured joints) and map it back to joint targets with a numerical damped-least-squares inverse-kinematics solver (warm-started from the current joints); the robot then executes the resulting joint chunk under real-time joint-position control. This conversion is performed entirely server-side, so the joint-space robot client is unchanged. Actions are executed with a near-synchronous chunking scheme: each 10-step delta-EEF chunk is executed in full at 5 Hz before the next inference, with a joint-velocity cap of 2.5 rad/s.

Comparison and protocol. Under identical deployment conditions we compare three policies: $\pi_{0.5}$, ACoT-VLA (the prior baseline), and our method (with latent spatiotemporal reasoning internalized through Scene Gist Memory). All three share the same fine-tuning configuration (delta-EEF action space, action horizon 10, `pi05_base` initialization, batch size 16, constant learning rate 5×10^{-5}) and differ only in the method itself. For each policy we run 16 trials with randomized initial screwdriver placements drawn from the same distribution across methods, and report the binary task-success rate.