SG-TTT: SCENE-GROUNDED SUPERVISION FOR TEST-TIME TRAINING OF WORLD ACTION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

World action models (WAMs), which couple action generation with predictive world modeling, can degrade substantially under distribution shifts between training and deployment. Test-time training (TTT) offers a way to adapt these models using observations collected during deployment. For WAM adaptation under deployment shifts, future-prediction TTT typically matches each predicted future signal with its realized counterpart, without explicitly exploiting relationships among signals that describe the same realized future state, potentially limiting the informativeness of supervision. We formulate supervision for future-prediction TTT as a two-step process: organizing observed signals and their relationships into a structured supervisory reference, then translating this reference into adaptation objectives. We propose **Scene-Grounded Test-Time Training (SG-TTT)**, a TTT method for WAMs with supervision grounded in scene geometry. SG-TTT first constructs a *factual scene anchor* from multi-view visual observations and the corresponding end-effector pose obtained after action execution, together with their spatial relationships. It then applies *component-wise scene consistency* to measure the scene-level spatial inconsistency introduced by each predicted future-state component and derives component-specific adaptation signals from the shared anchor. Extensive experiments on long-horizon LIBERO-10 suite under both standard and LIBERO-Plus distribution-shift settings, and real-world manipulation tasks show that SG-TTT achieves stronger overall performance than the frozen base WAM and representative future-prediction TTT baselines.

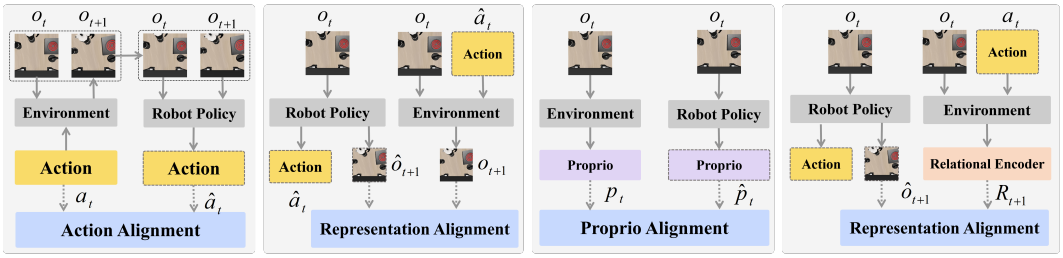
1 INTRODUCTION

World action models (WAMs) couple action generation with predictive world modeling in robot manipulation (Li et al., 2026; Kim et al., 2026). Despite their outstanding performance in robot manipulation, WAMs can suffer substantial degradation under deployment-time distribution shifts (Fei et al., 2026; Yang et al., 2023; Pang et al., 2025). Test-time training (TTT) offers a potential avenue for adapting WAMs using unlabeled deployment data, without requiring task labels or reward signals (Sun et al., 2019; Hansen et al., 2020).

Within this deployment-shift setting, Fig. 1 summarizes three common supervision interfaces for robot-policy TTT and contrasts them with our structured formulation of future-prediction supervision. Inverse-dynamics supervision reconstructs executed actions from observed transitions (Hansen et al., 2020; Yoneda et al., 2021), which is not naturally aligned with joint-modeling WAMs that generate actions and future states together (Kim et al., 2026; Ye et al., 2026). Current-state supervision predicts robot state from current visual observations (Zhang et al., 2026) and generally requires an auxiliary prediction path for future-predictive WAMs. Future-prediction supervision instead uses realized future observations or states to supervise quantities already predicted by the WAM (Yang et al., 2023; Wan et al., 2026; Park et al., 2026; Ji et al., 2026), making it the most directly aligned paradigm for WAM adaptation.

Effective TTT depends on the informativeness of its test-time supervision (Sun et al., 2019; Liu et al., 2021; Lu & Nemati, 2025). Despite this interface-level alignment, existing future-prediction TTT methods typically match each predicted future signal with its realized counterpart (Yang et al., 2023; Wan et al., 2026; Park et al., 2026; Ji et al., 2026), without explicitly exploiting relationships among signals that describe the same realized future state. We therefore formulate future-prediction

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.



(a) Inverse-Dynamics Supervision (b) Future-Prediction Supervision (c) Current-State Supervision (d) Structured Supervision (Ours)

Figure 1: **Representative supervision interfaces for adapting robot policies to deployment-time distribution shifts, together with our structured formulation of future-prediction supervision.** (a) Inverse-dynamics supervision; (b) direct future-prediction supervision; (c) current-state supervision; (d) structured future-prediction supervision (ours). o_t and o_{t+1} denote observations before and after an action; a_t, p_t, and R_{t+1} denote the executed action, current proprioceptive state, and structured supervisory reference, respectively. Hats and dashed boxes indicate predictions.

supervision for WAMs as a two-step process: organizing observed signals and their relationships into a structured supervisory reference, then translating this reference into adaptation objectives.

In robotic deployment, multi-view visual observations and robot end-effector pose provide complementary signals of the same scene, with their relationships naturally expressed through scene geometry. We instantiate this idea as Scene-Grounded Test-Time Training (SG-TTT). SG-TTT constructs a factual scene anchor from realized multi-view observations and the corresponding end-effector pose, and then applies component-wise scene consistency to measure the scene-level inconsistency introduced by each predicted future-state component relative to the shared anchor. Together, these mechanisms separate the factual evidence defining the supervisory context from the individual predicted component optimized in each routed branch, allowing heterogeneous deployment-time observations to constrain each prediction through their scene-level relationships.

Our contributions are threefold:

- We formulate future-prediction test-time supervision for WAMs as a two-step process: organizing observed signals and their relationships into a structured supervisory reference, then translating this reference into adaptation objectives for the WAM.
- We propose **Scene-Grounded Test-Time Training (SG-TTT)**, which instantiates this formulation through a factual scene anchor and component-wise scene consistency.
- We evaluate SG-TTT on long-horizon LIBERO-10 tasks under both standard and LIBERO-Plus distribution-shift settings, real-world manipulation tasks, together with representative future-prediction TTT baselines, ablations, and mechanism diagnostics.

2 Related Work

World-action models for robot control. World-action models (WAMs) couple robot action generation with predictive modeling of future world evolution. Video-predictive policies exploit future visual representations or jointly learned motion latents as intermediate representations for control (Hu et al., 2024; Routray et al., 2025). CauVA further unifies frame prediction and action inference within a single causal sequence, rather than following a strict imagine-then-execute pipeline (Li et al., 2026). Recent WAMs more tightly couple future-state and action generation through joint or coupled generative architectures (Kim et al., 2026; Ye et al., 2026; Zhou et al., 2026). Recent approaches also model future dynamics in compact latent or future-state representations, reducing dependence on dense pixel-space prediction (Lyu et al., 2026; Lei et al., 2026), while recent work also aligns world-model features with action-relevant representations at the world–action interface (Qiu et al., 2026). Across these designs, the main variations lie in architecture, prediction space, and the world–action interface. Our work is complementary: rather than redesigning the WAM architecture

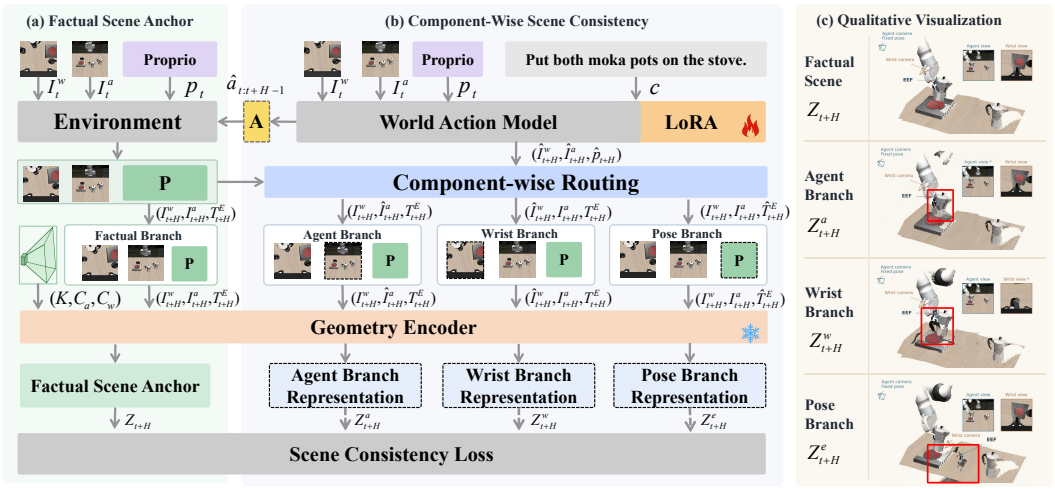


Figure 2: **Overview of SG-TTT.** The WAM predicts action chunk $\hat{a}_{t:t+H-1}$ and future components at $t+H$. (a) After execution, realized future images and EEF pose form the factual scene anchor through calibrated geometry. (b) Routing substitutes one predicted component; the shared geometry encoder produces joint representations for scene consistency supervision. A and P denote action and EEF pose; dashed component borders mark predictions. The snowflake and flame denote frozen and trainable modules. (c) Red boxes in scene visualizations highlight spatial inconsistencies caused by predicted agent-view, wrist-view, and pose components.

or prediction space, we study how deployment-time evidence can be organized into supervision for test-time adaptation of an existing WAM.

Test-time training for robot policies under deployment shifts. We focus on TTT methods that adapt robot policies to deployment-time distribution shifts using supervision derived from deployment-time evidence. Within this scope, representative supervisory objectives can be broadly organized according to the signals used as targets or references. Inverse-dynamics supervision predicts the intervening action from observations before and after a transition, using the action actually executed by the robot as the target (Hansen et al., 2020; Yoneda et al., 2021). Future-prediction supervision predicts future observations or representations from current observations and action-related information, using the realized future observations or their representations as targets (Yang et al., 2023; Wan et al., 2026; Park et al., 2026; Ji et al., 2026). Current-state supervision predicts the state of the robot from current visual observations, using robot measurements from the same time step as targets (Zhang et al., 2026). Among these paradigms, future-prediction supervision is most directly aligned with WAMs because it operates on quantities already exposed by their predictive interface. However, existing future-prediction TTT methods primarily use realized future signals as individual targets, without explicitly exploiting relationships among signals from the same realized future state as supervisory constraints. In contrast, SG-TTT explicitly organizes observed signals and their spatial relationships into a structured factual scene anchor, then translates this reference into component-specific adaptation objectives.

3 Method

3.1 Deployment Setting

SG-TTT first constructs a factual scene anchor from multi-view visual observations and the corresponding end-effector pose obtained after action execution, together with their spatial relationships. It then applies component-wise scene consistency to measure the scene-level spatial inconsistency introduced by each predicted future-state component, deriving component-specific adaptation signals for the WAM from the factual scene anchor.

Deployment-time evidence collection. Let $\mathcal{G}_{\theta,\psi}$ denote a future-predictive WAM, where θ remains fixed during deployment and ψ is adapted by TTT. At time step t, the context $\chi_t = (I_t^w, I_t^a, p_t, c)$ comprises the wrist-view image, agent-view image, proprioception, and task instruction. For an action horizon H, the WAM predicts an action chunk and future scene components:

$$(\hat{a}_{t:t+H-1}, \hat{y}_{t+H}) = \mathcal{G}_{\theta,\psi}(\chi_t), \tag{1}$$

where $\hat{y}_{t+H} = (\hat{I}^w_{t+H}, \hat{I}^a_{t+H}, \hat{T}^E_{t+H})$. For WAMs that predict proprioception $\hat{p}_{t+H}$, this interface includes the pose map $\hat{T}^E_{t+H} = \Pi_E(\hat{p}_{t+H})$ to obtain the end-effector (EEF) pose. During evidence collection, ψ is fixed at its initial value ψ_0.

Realized future observations. Executing $\hat{a}_{t:t+H-1}$ yields the realized multi-view visual observations and the EEF pose derived from proprioception. The resulting adaptation set of N_{TTT} samples is

$$\begin{aligned} y_{t+H} &= \left(I^w_{t+H}, I^a_{t+H}, T^E_{t+H}\right), \\ \mathcal{D}_{\mathrm{TTT}} &= \{(\chi_{t_i}, y_{t_i+H})\}_{i=1}^{N_{\mathrm{TTT}}}. \end{aligned} \tag{2}$$

This dataset remains fixed during adaptation. The realized future views and EEF pose provide the inputs for constructing the factual scene anchor in Fig. 2(a).

3.2 Factual Scene Anchor

The factual scene anchor jointly represents the realized future scene and serves as a common reference for all predicted components. Initial deployment calibration provides camera intrinsics $\mathbf{K} = (\mathbf{K}_a, \mathbf{K}_w)$ and reference extrinsics C_a, C_w. Given the realized future views and EEF pose, the frozen geometry encoder Φ_{geo} computes the corresponding camera geometry and jointly encodes both views conditioned on this geometry:

$$Z_{t+H} = \Phi_{\mathrm{geo}}\left(I^w_{t+H}, I^a_{t+H}, T^E_{t+H}; \mathbf{K}, C_a, C_w\right). \tag{3}$$

The factual scene anchor Z_{t+H} thus represents the two views together with their pose-dependent spatial relationships. The factual anchor and all routed scenes use the same encoder and calibration; coordinate normalization is detailed in Appendix B.

3.3 Component-Wise Scene Consistency

To obtain component-specific adaptation signals, *component-wise routing* replaces one realized future component with its prediction while keeping the remaining components fixed. This forms the agent, wrist, and EEF-pose branches in Fig. 2(b). Applying the shared geometry encoder gives

$$\begin{aligned} Z^a_{t+H} &= \Phi_{\mathrm{geo}}\Big(I^w_{t+H}, \hat{I}^a_{t+H}, T^E_{t+H}; \mathbf{K}, C_a, C_w\Big), \\ Z^w_{t+H} &= \Phi_{\mathrm{geo}}\Big(\hat{I}^w_{t+H}, I^a_{t+H}, T^E_{t+H}; \mathbf{K}, C_a, C_w\Big), \\ Z^e_{t+H} &= \Phi_{\mathrm{geo}}\Big(I^w_{t+H}, I^a_{t+H}, \hat{T}^E_{t+H}; \mathbf{K}, C_a, C_w\Big). \end{aligned} \tag{4}$$

Here Z^a_{t+H}, Z^w_{t+H}, and Z^e_{t+H} are joint feature representations of the three branches. As illustrated in Fig. 2(c), a predicted view can conflict with the other realized view and pose, whereas a predicted pose changes the wrist-camera geometry with both images fixed.

The *scene consistency loss* compares each routed branch with the same factual scene anchor:

$$\mathcal{L}_s = \mathcal{D}\left(Z^s_{t+H}, Z_{t+H}\right), \qquad s \in \{a, w, e\}, \tag{5}$$

where $\mathcal{D}$ measures the discrepancy between matched spatial features of the routed scene and factual anchor over both views. Its token-level formulation is given in Appendix B. Each loss therefore measures the scene-level spatial inconsistency introduced by its predicted component. Gradients pass through that component to ψ; the factual inputs and geometry encoder remain fixed.

Algorithm 1 Scene-Grounded Test-Time Training

Input: pretrained WAM $\mathcal{G}_{\theta,\psi_0}$; frozen geometry encoder Φ_{geo}; calibration $\mathbf{K}, C_a, C_w$; environment $\mathcal{E}$; sample budget N_{TTT}; epochs E_{TTT}; loss weights $(\lambda_a, \lambda_w, \lambda_e)$
Output: adapted WAM $\mathcal{G}_{\theta,\psi^\star}$

Step 1: Collect realized future observations
Collect N_{TTT} samples into $\mathcal{D}_{\text{TTT}}$ with the frozen WAM in $\mathcal{E}$ (Sec. 3.1)
Step 2: Construct factual scene anchors
for $(\chi_t, y_{t+H}) \in \mathcal{D}_{\text{TTT}}$ **do**
  $Z_{t+H} \leftarrow \Phi_{\text{geo}}(I^w_{t+H}, I^a_{t+H}, T^E_{t+H}; \mathbf{K}, C_a, C_w)$
end for
Step 3: Adapt with component-wise scene consistency
$\psi \leftarrow \psi_0$
for each of E_{TTT} epochs **do**
  for $(\chi_t, y_{t+H}) \in \mathcal{D}_{\text{TTT}}$ **do**
    Predict $\hat{y}_{t+H}$ with $\mathcal{G}_{\theta,\psi}(\chi_t)$ (Sec. 3.1)
    for $s \in \{a, w, e\}$ **do**
      Replace only component s of y_{t+H} with its prediction; keep the remaining components factual
      Encode the resulting scene with Φ_{geo} to obtain Z^s_{t+H} (Eq. 4)
      $\mathcal{L}_s \leftarrow \mathcal{D}(Z^s_{t+H}, Z_{t+H})$
    end for
    $\mathcal{L}_{\text{SG}} \leftarrow \lambda_a \mathcal{L}_a + \lambda_w \mathcal{L}_w + \lambda_e \mathcal{L}_e$
    Update only ψ using $\nabla_\psi \mathcal{L}_{\text{SG}}$
  end for
end for
$\psi^\star \leftarrow \psi$; freeze $\psi^\star$
return $\mathcal{G}_{\theta,\psi^\star}$

3.4 Scene-Grounded Test-Time Training Framework

The scene-grounded objective combines the three component-specific losses:

$$\boxed{\mathcal{L}_{\text{SG}} = \lambda_a \mathcal{L}_a + \lambda_w \mathcal{L}_w + \lambda_e \mathcal{L}_e.} \tag{6}$$

The nonnegative weights $\lambda_a, \lambda_w, \lambda_e$ correspond to the agent, wrist, and EEF-pose branches, respectively. Starting from ψ_0, SG-TTT optimizes ψ on the fixed adaptation dataset:

$$\psi^\star = \arg\min_{\psi} \frac{1}{|\mathcal{D}_{\text{TTT}}|} \sum_{(\chi_t, y_{t+H}) \in \mathcal{D}_{\text{TTT}}} \mathcal{L}_{\text{SG}}(\hat{y}_{t+H}(\psi), y_{t+H}), \qquad \theta \text{ fixed.} \tag{7}$$

At each update, the WAM regenerates future predictions using the current ψ. Supervision on future components updates ψ, which also affects action generation (Appendix A). Algorithm 1 summarizes adaptation over E_{TTT} epochs. Implementation and optimization details are given in Appendices B and C.3.

After adaptation, $\psi^\star$ is frozen and the supervision modules are discarded. The adapted WAM is evaluated on disjoint episodes without further updates.

4 Experiments

4.1 Experimental Setup

Benchmarks. We evaluate SG-TTT on the long-horizon LIBERO-10 suite from both LIBERO (Liu et al., 2023) and LIBERO-Plus (Fei et al., 2026). LIBERO provides the nominal setting, while LIBERO-Plus introduces seven deployment shifts: object layout, camera viewpoint, robot initialization, lighting, background texture, sensor noise, and language variation. We further evaluate SG-TTT on three real-world manipulation tasks using the reBot Arm B601-DM (Seeed Studio) under nominal and shifted conditions.

Baselines. We compare against frozen Cosmos Policy (Kim et al., 2026) and three direct-matching future-prediction TTT baselines (Wan et al., 2026): Visual-only TTT optimizes future image prediction, Proprio-only TTT optimizes future proprioceptive-state prediction, and Direct-future TTT combines both objectives independently.

Implementation and Evaluation Details. We instantiate the WAM with Cosmos Policy and adapt it using LoRA (Hu et al., 2021). The geometry encoder uses frozen DA3 (Lin et al., 2025) for joint spatial feature extraction. In simulation, all methods share the base checkpoint; adaptive methods use 64 unlabeled TTT trajectory samples and 64 optimizer updates. We report mean task success rates over three independent adaptations for LIBERO, LIBERO-Plus, and all simulation ablations. Optimization, evaluation, and statistical details are provided in Appendix C.

4.2 Main Results

We evaluate SG-TTT on the LIBERO-10 suite under both standard LIBERO and LIBERO-Plus distribution-shift settings, the latter covering seven controlled perturbation dimensions.

4.2.1 Nominal Setting

Table 1 compares different choices of test-time supervision on nominal LIBERO-10. Visual-only and Proprio-only TTT both improve over the frozen WAM, showing that individual deployment signals provide useful adaptation evidence. However, Direct-future TTT combines visual and proprioceptive prediction objectives through independent matching and does not improve upon the strongest single-signal baseline in mean success rate. This suggests that adding prediction targets alone does not guarantee more effective test-time supervision.

SG-TTT uses multi-view visual observations and EEF pose as complementary scene signals and organizes them through a structured scene reference. It achieves the highest mean success rate among the compared methods. Together, these comparisons motivate examining both the choice of deployment evidence and how heterogeneous signals are organized into adaptation objectives. Section 4.3 further examines this supervision design under controlled distribution shifts.

Table 1: Nominal LIBERO-10 success rates (S.R.); TTT results are mean $\pm$ SD over three adaptations.

Method	TTT evidence	Supervision construction	S.R. (%) $\uparrow$
No adaptation			
Cosmos Policy	–	–	97.20
Direct-matching TTT			
Visual-only TTT	Visual	Direct matching	98.00 ± 0.43
Proprio-only TTT	Proprioception	Direct matching	98.27 ± 0.19
Direct-future TTT	Visual + Proprioception	Independent matching	98.07 ± 0.82
Proposed method			
SG-TTT	Multi-view visual + EEF pose	**Structured scene reference**	$\mathbf{98.87 \pm 0.09}$

4.2.2 Robustness Under Distribution Shifts

Table 2 reports results on the LIBERO-10 suite of LIBERO-Plus (Fei et al., 2026) across seven perturbation types. SG-TTT achieves the highest average success rate of 82.5%, improving over frozen Cosmos Policy by 1.8 percentage points and Direct-future TTT by 1.4 points. It outperforms Direct-future TTT on all seven perturbation types and leads or ties for the best result on five. The gains vary across shifts, and the relative performance of the direct-matching baselines also changes.

Visual-only TTT improves performance under lighting and sensor noise, but falls below the frozen WAM under background changes. Its image-matching objective thus yields different outcomes even among appearance perturbations. Proprio-only TTT achieves the highest success rate under camera shifts, yet reduces performance under lighting changes. Robot-state supervision is useful, although its objective does not explicitly constrain the corresponding visual scene. Direct-future TTT combines these two prediction objectives independently, but ranks below both single-signal baselines in average success and falls below the frozen WAM on layout shifts. Having both types of evidence therefore does not guarantee that their independent prediction losses produce better adaptation.

SG-TTT obtains its largest gain over the frozen WAM under robot initialization, which perturbs initial joint positions. Camera shifts change camera distance, position, or orientation. These perturbations

Table 2: LIBERO-Plus under seven official perturbation dimensions. Cells report mean episode success rates (%) over three independent adaptations. Avg. equally weights the seven perturbation dimensions.

Method	Layout	Camera	Robot Init	Light	Background	Noise	Language	Avg.
No adaptation								
Cosmos Policy	92.3%	58.9%	62.6%	96.0%	71.3%	86.6%	**97.1%**	80.7%
Direct-matching TTT								
Visual-only TTT	93.3%	61.3%	65.1%	**97.4%**	70.2%	**87.3%**	95.3%	81.4%
Proprio-only TTT	92.9%	**62.1%**	64.9%	93.4%	72.3%	86.2%	96.3%	81.2%
Direct-future TTT	91.7%	60.4%	64.6%	97.1%	71.6%	87.1%	95.6%	81.1%
Proposed method								
SG-TTT	**94.2%**	61.8%	**66.4%**	**97.4%**	**73.7%**	**87.3%**	96.9%	**82.5%**

alter the correspondence among deployment-time signals, whose relationships are encoded by the factual scene anchor as a shared supervisory reference. SG-TTT improves over the frozen WAM under camera shifts, but Proprio-only TTT remains slightly stronger. The benefit of structured supervision relative to single-signal objectives therefore varies with the shift. SG-TTT also leads under layout changes, which add distractors or perturb target-object poses. The factual scene anchor provides a reference for the resulting deployment scene.

SG-TTT also achieves the best result under background variation and ties for the best under lighting changes. These perturbations primarily affect appearance while preserving scene structure, suggesting that relational scene constraints remain useful beyond geometric changes. Under language variation, SG-TTT remains close to the frozen WAM without improving on it; its objective provides no direct language-alignment supervision.

These results suggest that effective test-time supervision depends on both the available signals and how their relationships are organized into adaptation objectives. Section 4.3 further analyzes the roles of the factual scene anchor and component-wise routing under matched controls.

4.3 Analysis of Relational Supervision Design

Although SG-TTT consistently improves adaptation under diverse deployment shifts, the source of this improvement is not merely the availability of additional prediction signals. We therefore analyze whether heterogeneous deployment observations should be optimized independently or organized through relational supervision. Table 3 progressively introduces relational reference construction and then component-wise attribution. The first stage tests the effect of organizing heterogeneous signals through a shared factual scene reference, while the second stage examines how the resulting discrepancies should be assigned to individual prediction components. All adaptive variants use matched adaptation data and update budgets.

4.3.1 Two-Step Relational Supervision Design

Shared relational reference. Independent-component supervision provides complementary adaptation signals, but each prediction branch is optimized against its own discrepancy without explicitly constraining their correspondence. Introducing a shared factual scene reference improves the average success from 65.6% to 66.4%, with consistent gains across all three shifts. Its role is to provide a common relational constraint rather than an additional prediction target. This indicates that the benefit does not originate from adding more prediction targets alone; instead, organizing heterogeneous observations through a common reference provides a more reliable adaptation signal.

Component-wise relational attribution. Building upon the shared reference, component-wise routing further improves the average success to 67.3%. By activating one predicted component while keeping the remaining components factual, each branch measures the relational discrepancy associated with its own prediction path. The improvement is particularly evident under background variations (+2.1% over the joint-reference variant), where disentangling component-specific inconsistencies becomes more important. These results support the two-step design: a shared reference

Table 3: Relational supervision design and branch contribution analysis. Entries are success rates (%) averaged over three independent adaptations; Avg. (3) equally weights the three dimensions. The visual scene reference row provides a visual-only reference control without EEF-pose input, while the branch analysis separately evaluates the contribution of EEF-pose information within the full relational supervision framework. Bold marks column-wise best results within each block, including ties.

Configuration	Camera	Robot init.	Background	Avg. (3)↑
Staged supervision design				
Independent components	60.9%	64.6%	71.3%	65.6%
+ Shared factual reference	61.6%	65.9%	71.6%	66.4%
+ Component-wise routing	**61.8%**	**66.4%**	**73.7%**	**67.3%**
Scene evidence and branch supervision				
Full (all branches)	61.8%	**66.4%**	**73.7%**	**67.3%**
− Agent-view branch	62.1%	64.6%	73.4%	66.7%
− Wrist-view branch	62.1%	65.1%	72.0%	66.4%
− EEF-pose branch	**62.5%**	65.6%	73.4%	67.2%
Visual scene reference	61.3%	65.1%	71.3%	66.0%

establishes cross-signal relational consistency, while routing assigns the resulting supervision to the corresponding prediction component.

4.3.2 Contribution of Different Relational Cues

We further analyze the role of each branch in Table 3. Removing the wrist-view, agent-view, or EEF-pose branch leads to different trade-offs across shifts rather than uniformly degrading every dimension. In particular, the EEF-pose branch has a shift-dependent contribution: removing it slightly improves camera performance while reducing adaptation under robot initialization and background changes. In contrast, the EEF-pose information remains consistently useful when used as a relational cue within SG-TTT. This suggests that pose information is more useful when incorporated as a relational cue among deployment-time signals than as an isolated prediction objective. Together, these observations highlight that SG-TTT benefits from how signals are organized into relational supervision, rather than from any single modality alone.

4.4 Real-World Experiments

To further evaluate whether relational supervision remains effective under physical deployment conditions, we conduct real-world experiments on the B601-DM robot platform. Following the observation shifts studied in LIBERO-Plus (Fei et al., 2026), we consider two physical counterparts in deployment: camera variation and robot initialization variation (Fig. 4). We evaluate SG-TTT on three manipulation tasks with different interaction complexity, ranging from short-horizon object manipulation to precise and long-horizon execution (Fig. 3).

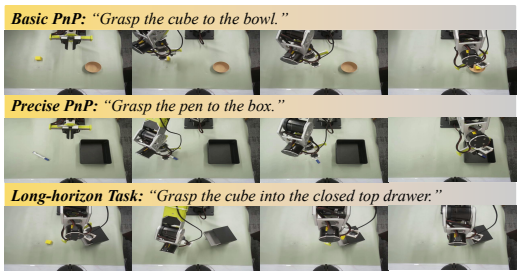


Figure 3: **Real-world task examples.** Each row shows a task sequence from left to right. PnP denotes pick-and-place.

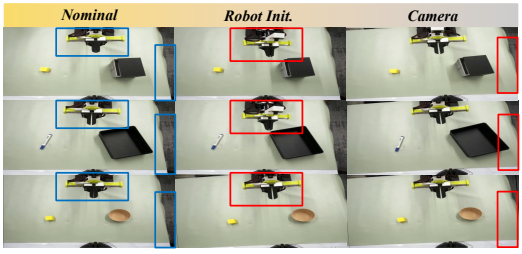


Figure 4: **Real-world evaluation conditions.** Blue boxes mark nominal references; red boxes highlight changes under Robot Init. and Camera.

For SG-TTT, the frozen Cosmos Policy first collects a small unlabeled interaction set for each task-condition. We perform adaptation once on this fixed set and freeze the adapted checkpoint for all subsequent evaluation rollouts.

Table 4: B601-DM real-world results. Score and SR are reported as percentages; averages equally weight the three tasks.

		Cube → Bowl		Pen → Box		Cube → Top Drawer			
Method	Scenario	Score	SR	Score	SR	Score	SR	Avg. Score	Avg. SR
GR00T N1.7-3B	Nominal	45.0%	40.0%	70.0%	70.0%	55.0%	40.0%	56.7%	50.0%
	Robot Init	20.0%	20.0%	55.0%	50.0%	40.0%	20.0%	38.3%	30.0%
	Camera	20.0%	20.0%	40.0%	40.0%	12.5%	10.0%	24.2%	23.3%
Cosmos Policy	Nominal	60.0%	60.0%	50.0%	50.0%	85.0%	80.0%	65.0%	63.3%
	Robot Init	45.0%	40.0%	35.0%	30.0%	60.0%	50.0%	46.7%	40.0%
	Camera	35.0%	30.0%	30.0%	20.0%	47.5%	40.0%	37.5%	30.0%
Cosmos + SG-TTT	Nominal	70.0%	70.0%	65.0%	60.0%	90.0%	90.0%	75.0%	73.3%
	Robot Init	70.0%	70.0%	45.0%	40.0%	67.5%	60.0%	60.8%	56.7%
	Camera	55.0%	50.0%	45.0%	40.0%	57.5%	50.0%	52.5%	46.7%

Table 4 summarizes the real-world adaptation results under both nominal and shifted settings. Compared with the frozen Cosmos Policy, SG-TTT consistently improves deployment performance across physical conditions, indicating that unlabeled interaction data can provide effective adaptation supervision beyond simulation environments. The improvement under Robot Init variation highlights the role of relational supervision in maintaining correspondence between visual observations and robot states when the physical configuration changes. Under Camera variation, SG-TTT preserves robust performance despite changes in visual observations, suggesting that the relational reference remains informative beyond the original deployment view.

We additionally report GR00T N1.7-3B (NVIDIA et al., 2025) as an external system-level reference. Since it is not a matched test-time adaptation baseline, we use it only as a broader performance reference rather than a direct comparison of adaptation capability. Overall, these results show that relational supervision constructed from real robot interactions remains effective beyond simulation, supporting the applicability of relational supervision for adapting world action models under physical deployment shifts.

We report **Score**, the normalized fraction of ordered stages completed (Pick-Place for the first two tasks; Open-Pick-Place-Close for the drawer), and **success rate (SR)**, the fraction of fully successful rollouts. Appendix E gives the exact task, reset, metric, and shift protocols.

5 Conclusion

We presented SG-TTT, a test-time training method for WAMs that constructs relational supervision from deployment-time signals through scene geometry. Its factual scene anchor organizes multi-view observations, the corresponding end-effector pose, and their spatial relationships into a shared supervisory reference. Component-wise scene consistency assesses the scene-level spatial inconsistency introduced by individual predicted future-state components relative to this reference, yielding component-specific adaptation signals. Experiments on LIBERO-10 under standard and LIBERO-Plus distribution-shift settings, and on real-world manipulation tasks, demonstrate that SG-TTT achieves stronger overall performance than the frozen base WAM and representative future-prediction TTT baselines. More broadly, SG-TTT suggests a design principle for future-prediction TTT of WAMs under deployment-time distribution shifts: exploiting relationships among deployment-time signals can complement supervision on individual future predictions, opening a broader design space for informative adaptation objectives derived from unlabeled interaction data.

AI USE STATEMENT

We used generative AI tools to assist with drafting and editing the manuscript, literature retrieval and organization, research ideation and experimental design, method implementation, and the gradient derivations in Appendix A. These tools were not used to generate synthetic datasets or experimental results. The authors reviewed the AI-assisted material and take full responsibility for the final manuscript and associated code, including the mathematical arguments and reported results.

ETHICS STATEMENT

This work studies test-time adaptation for robotic manipulation in simulation and on a physical robot. We adhere to the ICLR Code of Ethics. The real-robot evaluation protocol includes joint and speed limits and safety-based termination, with human intervention permitted for emergency stopping. Improved task performance does not constitute a guarantee of safe operation outside the evaluated settings. Deployment in other environments requires application-specific safety validation and appropriate safeguards.

REPRODUCIBILITY STATEMENT

Section 3 and Algorithm 1 describe the SG-TTT objectives and adaptation procedure. Appendix B provides implementation details for scene construction and component-wise consistency. Appendix C describes the adaptation and evaluation splits, optimization settings, and evaluation metrics. The real-robot platform, data collection, and evaluation protocol are described in Appendix E. Adaptation and evaluation use disjoint records, and controlled comparisons use matched data and optimization budgets.

REFERENCES

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 (CVPR)*, pp. 38574–38583, June 2026.

Nicklas Hansen, Rishabh Jangir, Yu Sun, Guillem Alenyà, P. Abbeel, Alexei A. Efros, Lerrel Pinto, and Xiaolong Wang. Self-supervised policy adaptation during deployment. *ArXiv*, abs/2007.04309, 2020. URL https://api.semanticscholar.org/CorpusID:220404330.

J. Edward Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, She-An Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *ArXiv*, abs/2106.09685, 2021. URL https://api.semanticscholar.org/CorpusID:235458009.

Yucheng Hu, Yanjiang Guo, Pengchao Wang, Xiaoyu Chen, Yen-Jen Wang, Jianke Zhang, Koushil Sreenath, Chaochao Lu, and Jianyu Chen. Video prediction policy: A generalist robot policy with predictive visual representations. *ArXiv*, abs/2412.14803, 2024. URL https://api.semanticscholar.org/CorpusID:274859626.

Hong Chang Ji, Guo-Yang Xia, Luo-Yang Sun, Fangxiang Feng, and Lei Ren. Vane: Reliable test-time training for vision-language-action models via future visual representation prediction. 2026. URL https://api.semanticscholar.org/CorpusID:290991059.

Moo Jin Kim, Yihuai Gao, Tsung-Yi Lin, Yen-Chen Lin, Yunhao Ge, Grace Tin Yan Lam, Percy Liang, Shuran Song, Ming-Yu Liu, Chelsea Finn, and Jinwei Gu. Cosmos policy: Fine-tuning video models for visuomotor control and planning. *ArXiv*, abs/2601.16163, 2026. URL https://api.semanticscholar.org/CorpusID:284935587.

Fenghao Lei, Zhixiong Huang, Long Yang, Jia-Bao Chen, Pei-Lin Huang, Han Fu, Zhuo Li, and Xiaoxue Ren. Dele-w0.5: Inferring action from future latent state for robotic manipulation. 2026. URL https://api.semanticscholar.org/CorpusID:291324579.

Lin Li, Qihang Zhang, Yiming Luo, Shuai Yang, Ruilin Wang, Luyao Zhang, Mingrui Yu, Zelin Gao, Nan Xue, Boyu Zhou, Xing Zhu, Mingyu Ding, Yujun Shen, and Yinghao Xu. Causal World Modeling for Robot Control. In *Proceedings of Robotics: Science and Systems*, Sydney, Australia, July 2026. doi: 10.15607/RSS.2026.XXII.016.

Haotong Lin, Si-Lei Chen, Jun Hao Liew, Donny Y. Chen, Zhenyu Li, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth anything 3: Recovering the visual space from any views. *ArXiv*, abs/2511.10647, 2025. URL https://api.semanticscholar.org/CorpusID:282992334.

Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qiang Liu, Yuke Zhu, and Peter Stone. Libero: Benchmarking knowledge transfer for lifelong robot learning. *ArXiv*, abs/2306.03310, 2023. URL https://api.semanticscholar.org/CorpusID:259089508.

Yuejiang Liu, Parth Kothari, Bastien van Delft, Baptiste Bellot-Gurlet, Taylor Mordan, and Alexandre Alahi. Ttt++: When does self-supervised test-time training fail or thrive? In *Neural Information Processing Systems*, 2021. URL https://api.semanticscholar.org/CorpusID:247410579.

Xiao-Lei Lu and Shamim Nemati. Adaptive test-time training for predicting need for invasive mechanical ventilation in multi-center cohorts. *ArXiv*, abs/2512.06652, 2025. URL https://api.semanticscholar.org/CorpusID:283693608.

Jiangran Lyu, Kai Liu, Xuheng Zhang, Hao-Ran Liao, Yusen Feng, Wenxuan Zhu, Tingrui Shen, Jiayi Chen, Jiazhao Zhang, Yifei Dong, Wen-Bo Cui, Senmao Qi, Shuo Wang, Yi-Xin Zheng, Mingqi Yan, Xuesong Shi, Haoran Li, Dong-Bin Zhao, Mingyu Liu, Zhizheng Zhang, Li Yi, Yizhou Wang, and He Wang. Lda-1b: Scaling latent dynamics action model via universal embodied data ingestion. *ArXiv*, abs/2602.12215, 2026. URL https://api.semanticscholar.org/CorpusID:285540999.

NVIDIA, Johan Bjorck, Nikita Cherniadev Fernando Castañeda, Xingye Da, Runyu Ding, Linxi "Jim" Fan, Yu Fang, Dieter Fox, Fengyuan Hu, Spencer Huang, Joel Jang, Zhenyu Jiang, Jan Kautz, Kaushil Kundalia, Lawrence Lao, Zhiqi Li, Zongyu Lin, Kevin Lin, Guilin Liu, Edith Llontop, Loic Magne, Ajay Mandlekar, Avnish Narayan, Soroush Nasiriany, Scott Reed, You Liang Tan, Guanzhi Wang, Zu Wang, Jing Wang, Qi Wang, Jiannan Xiang, Yuqi Xie, Yinzhen Xu, Zhenjia Xu, Seonghyeon Ye, Zhiding Yu, Ao Zhang, Hao Zhang, Yizhou Zhao, Ruijie Zheng, and Yuke Zhu. GR00T N1: An open foundation model for generalist humanoid robots. In *ArXiv Preprint*, March 2025.

Jing-Cheng Pang, Nan Tang, Kaiyuan Li, Yuting Tang, Xin-Qiang Cai, Zhen-Yu Zhang, Gang Niu, Masashi Sugiyama, and Yang Yu. Learning view-invariant world models for visual robotic manipulation. In *International Conference on Learning Representations*, 2025. URL https://api.semanticscholar.org/CorpusID:280016080.

Sang Yoon Park, Wonjoong Kim, Yeonjun In, Sein Kim, Hongseok Kang, and Chanyoung Park. Test-time training for visual foresight vision-language-action models. *ArXiv*, abs/2605.08215, 2026. URL https://api.semanticscholar.org/CorpusID:288260229.

Lu Qiu, Yi-Zhuo Li, Yi Chen, Yu-Ying Ge, Yi-Xiao Ge, and Xihui Liu. Making foresight actionable: Repurposing representation alignment in world action models. *ArXiv*, abs/2606.12217, 2026. URL https://api.semanticscholar.org/CorpusID:289167329.

Sandeep Routray, Hengkai Pan, Unnat Jain, Shikhar Bahl, and Deepak Pathak. Vipra: Video prediction for robot actions. *ArXiv*, abs/2511.07732, 2025. URL https://api.semanticscholar.org/CorpusID:282922387.

Yu Sun, X. Wang, Zhuang Liu, John Miller, Alexei A. Efros, and Moritz Hardt. Test-time training with self-supervision for generalization under distribution shifts. In *International Conference on Machine Learning*, 2019. URL https://api.semanticscholar.org/CorpusID:220301705.

Chi Wan, Kangrui Wang, Yuan Si, Pingyue Zhang, and Manling Li. Worldagen: Unified state-action prediction with test-time world model training. In *AAAI Conference on Artificial Intelligence*, 2026. URL https://api.semanticscholar.org/CorpusID:286707669.

Sizhe Yang, Yanjie Ze, and Huazhe Xu. Movie: Visual model-based policy adaptation for view generalization. *ArXiv*, abs/2307.00972, 2023. URL https://api.semanticscholar.org/CorpusID:259317193.

Seonghyeon Ye, Yunhao Ge, Kai-Yuan Zheng, Shenyuan Gao, Sihyun Yu, George Kurian, Suneel Indupuru, You Tan, Chu-Ning Zhu, Jian-Nan Xiang, Aravind Malik, Kyungmin Lee, William Liang, Nadun Ranawaka, Jiasheng Gu, Yinzhen Xu, Guan-Zhi Wang, Fengyuan Hu, Avnish Narayan, Johan Bjorck, Jing Wang, Gwanghyun Kim, Dantong Niu, Ruijie Zheng, Yuqi Xie, Jimmy Wu, Qi Wang, Ryan C. Julian, Dan Xu, Yilun Du, Yevgen Chebotar, Scott Reed, Jan Kautz, Yuke Zhu, Linxi Jim Fan, and Joel Jang. World action models are zero-shot policies. *ArXiv*, abs/2602.15922, 2026. URL https://api.semanticscholar.org/CorpusID:285726165.

Takuma Yoneda, Yang Ge, Matthew R. Walter, and Bradly C. Stadie. Invariance through latent alignment. *Robotics: Science and Systems XVIII*, 2021. URL https://api.semanticscholar.org/CorpusID:248830814.

Wenbo Zhang, Jianxiong Li, Shuai Yang, Sijin Chen, Jiajun Liu, Lingqiao Liu, and Xiao Ma. Ttt-vla: Test-time latent prompt optimization for vision-language-action models. *ArXiv*, abs/2606.03127, 2026. URL https://api.semanticscholar.org/CorpusID:288863159.

Pengfei Zhou, Shengcong Chen, Di Chen, Jia-Xu Wang, Rong-Jun Jin, Bin Zhu, Yike Pan, Song-En Gu, Kuanning Wang, Shufeng Nan, Xingyu Qiu, Chen-Hao Qiu, Pu Yang, Yunuo Cai, Jianxiong Gao, Yifan Li, Yanwei Fu, Xiangyu Yue, Zhi Chen, and Jianlan Luo. τ0-wm: A unified video-action world model for robotic manipulation. *ArXiv*, abs/2606.01027, 2026. URL https://api.semanticscholar.org/CorpusID:288861329.