METAREACH: META-REACHABILITY GROUP RELATIVE POLICY OPTIMIZATION FOR VIDEO WORLD MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Conditional video world models predict future visual states under given actions, and post-training them with group relative policy optimization (GRPO) is essential for improving generation quality. However, existing methods face two challenges: (1) they freeze the decoder and reward against ground-truth frames, introducing an irreducible residual that distorts ranking due to the decoder's limited reachable space (target-set mismatch); and (2) they broadcast a single sequence-level advantage to all token blocks, ignoring the causal structure of autoregressive generation (long-horizon credit misallocation). To address these challenges, we propose MetaReach, a bi-level meta-learning framework. The inner loop adapts the decoder to reconstruct frames within the world model's output distribution, while the outer loop updates the world model under the adapted decoder. This reachability-calibrated meta-learning (RCM) ensures reward comparisons within the same reachable space, eliminating target-set mismatch. We introduce autoregressive credit assignment (ACA), which assigns a position-specific advantage to each token block according to its future horizon, aligning credit with autoregressive causality without auxiliary value networks or tree-structured rollouts. A gradient projection step preserves progress on the original objective. Experiments on RT-1, RoboNet, and VP2-RoboSuite over a 15-frame prediction horizon demonstrate that MetaReach consistently outperforms GRPO-based baselines and recent long-horizon post-training methods, reducing LPIPS by 7.7% and MSE by 13.3% over RLVR-World on RT-1.

1 INTRODUCTION

Conditional video world models are a core component of embodied intelligence, enabling agents to imagine future visual states under given actions and supporting planning, simulation, and decision-making (Goff et al., 2025; Zhao et al., 2025; Li et al., 2026). Recently, autoregressive architectures based on discrete tokens have become mainstream: videos are encoded into compact discrete tokens, the world model predicts future tokens in latent space, and a decoder reconstructs images (Wang et al., 2025a; Wu et al., 2025; Ren et al., 2025). This discrete token formulation facilitates autoregressive pretraining on large-scale video data, and its decoded outputs can be supervised by task-related rewards, thereby supporting post-training based on candidate comparison (Wu et al., 2025; Zhu et al., 2026). However, this paradigm faces two fundamental challenges: (1) *target-set mismatch* causes ranking distortion, and (2) *long-horizon credit misallocation* causes optimization bias. These issues jointly lead to unstable post-training and degraded long-horizon generation quality.

The first challenge stems from the reachable output space constraint. Quantization irreversibly discards information (Wang et al., 2025d), and the decoder capacity is limited, so it can only generate a restricted subset of all possible images, referred to as the "reachable output space" (Vafa et al., 2024). Real observations often contain details not preserved by quantization or out-of-distribution content, and typically lie outside this space (Albuquerque et al., 2024). Using real frames as reward references leaves an irreducible fixed residual for all candidates, which distorts ranking, i.e., *target-set mismatch* (Vafa et al., 2024; Wu et al., 2025). The second challenge stems from long-horizon credit misallocation. In autoregressive generation, early predictions condition all subsequent frames and thus have greater influence. However, existing GRPO-based methods broadcast sequence-level

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.

advantages uniformly to all future tokens, ignoring temporal asymmetry, and this bias accumulates over long horizons (Samsami et al., 2024; Savov et al., 2025). These two issues jointly constrain post-training effectiveness, motivating a revisit of evaluation targets and temporal credit assignment.

To address these challenges, existing methods fall into three categories. The first improves tokenization or representation learning to enhance latent space fidelity, but the frozen decoder's output space remains constrained, leaving real observations potentially unreachable (Wang et al., 2025b; Kim et al., 2026). The second addresses temporal credit assignment via sequence-level strategies (Cohen et al., 2024; Ramesh et al., 2024), but its coarse granularity misaligns with the causal influence of early predictions. The third augments training signals via reward shaping or auxiliary objectives (Hutson et al., 2024; Verma & Metcalf, 2024), but often requires task-specific priors and may introduce additional bias. Beyond these, Group Relative Policy Optimization (GRPO) is introduced into world model post-training, optimizing policies via group-relative comparisons among candidate trajectories (Wu et al., 2025; Zhu et al., 2026; Shao et al., 2024). However, existing GRPO-based methods exhibit two key limitations: (1) they rank candidates against real frames that often lie outside the reachable output space; (2) they broadcast sequence-level advantages to all future tokens, ignoring temporal asymmetry. Despite prior efforts, a unified framework that jointly addresses ranking distortion and long-horizon credit assignment remains absent.

To fill this gap, we propose MetaReach (Meta-Reachability Group Relative Policy Optimization), which redesigns post-training signals from two complementary perspectives: meta-learning-based evaluator calibration and temporal credit assignment. We cast post-training as a bi-level meta-learning optimization, rather than a static reward substitution. The inner loop updates the decoder to reconstruct the current frame from the input frame, turning the frozen autoencoder into an adaptive autoregressive reconstructor. The outer loop updates the world model to predict the next frame, meta-learning the parameters that maximize primary-task performance under the adapted decoder (Yao et al., 2025; Yoo et al., 2025). By alternating between the two loops, the decoder is dynamically calibrated to the reachable space of the world model's predictions, and the world model generates frames that are both accurate and decodable. This meta-learning calibration replaces the original real anchors with reachable reconstructions that adapt during training, eliminating *target-set mismatch* at the source. We further introduce an autoregressive credit assignment strategy that propagates per-frame rewards from each timestep to the sequence end and assigns them exclusively to their corresponding token blocks, aligning credit assignment with autoregressive dependencies.

The main contributions are summarized as follows:

- We propose MetaReach, a unified post-training framework that integrates meta-learning inner-outer loops for reachability calibration and structured temporal credit assignment in autoregressive video generation.
- We introduce the Reachability-Calibrated Meta-learning (RCM) module with inner-outer loops, which ensures candidate comparison within the reachable space and preserves the original optimization direction via gradient projection.
- We design the Autoregressive Credit Assignment (ACA) strategy, which distributes discounted returns to their corresponding token blocks, mitigating long-horizon credit misallocation by aligning credit with causal dependencies.
- Experiments on standard benchmarks demonstrate that MetaReach achieves state-of-the-art performance. Ablations validate the RCM module and the ACA strategy.

2 Related Work

2.1 World Models for Video Generation

World models are a foundational paradigm for generative video prediction (Feng et al., 2025; Ding et al., 2026). Early approaches employ recurrent networks and state-space models, while Transformers transform the field, with iVideoGPT (Wu et al., 2024), EMERALD (Burchi & Timofte, 2025), and Motus (Bi et al., 2026) achieving superior sample efficiency. Diffusion models offer a powerful alternative for high-fidelity generation, with Epona (Zhang et al., 2025) and Vid2World (Huang et al., 2026) succeeding in autonomous driving. Recent work explores object-centric decomposition (Nishimoto & Matsubara, 2026), fast-slow dynamics coordination (Wang et al., 2026a), unified

planning and generation (Xia et al., 2026), and 4D reconstruction from monocular videos (Yang et al., 2026b). Self-supervised learning dominates pre-training on large-scale video via next-token prediction or masked objectives (Zhang et al., 2023; Micheli et al., 2024; Burchi & Timofte, 2025; Wu et al., 2024), and reinforcement learning adapts these models for downstream tasks using rollouts (Wu et al., 2025; Zhu et al., 2026). However, post-training, which refines generated trajectories against target references, remains underexplored (Wang et al., 2025c; Dang et al., 2026). This gap motivates our review of post-training methods and their limitations.

2.2 Post-Training for World Models

Post-training has been explored along two directions corresponding to our targeted limitations. (1) *Evaluation target construction*: tokenization improvements enhance latent fidelity (Wu et al., 2025; Yan et al., 2025) but leave the decoder's reachable space constrained; reward shaping and auxiliary objectives (Verma & Metcalf, 2024; Rigter et al., 2024) require task-specific priors and risk bias; latent-space alignment avoids noisy pixel supervision (Jia et al., 2026); and GRPO-based post-training (Wu et al., 2025; Zhu et al., 2026), extended by WorldCompass with clip-level rollouts (Wang et al., 2026b), still ranks candidates against ground-truth frames outside the reachable space, causing *target-set mismatch* (Vafa et al., 2024; Dang et al., 2026). (2) *Temporal credit assignment*: sequence-level strategies (Cohen et al., 2024; Ramesh et al., 2024) rely on coarse-grained attribution; long-context modeling captures temporal dependencies (Po et al., 2025; Yang et al., 2026a), and Self Forcing eliminates exposure bias via self-rollout (Huang et al., 2025). However, they broadcast sequence-level advantages uniformly across all future tokens, neglecting autoregressive temporal asymmetry. Consequently, to date, a unified framework addressing both reachability-induced ranking distortion and long-horizon credit assignment remains absent, motivating our adoption of meta-learning in this work.

2.3 Meta-Learning and Bi-Level Optimization

To jointly address the two limitations above, we turn to meta-learning, which enables rapid adaptation by learning how to learn (Finn et al., 2017; Antoniou et al., 2019; Zintgraf et al., 2019). Its central formulation, bi-level optimization, employs an inner loop to adapt task-specific parameters and an outer loop to meta-learn hyperparameters or initialization, with recent advances handling lower-level constraints via regularized gap functions (Yao et al., 2025). This formulation is well suited when training and evaluation objectives are misaligned, as in world model post-training where decoder reconstruction and world model prediction need not coincide. In video generation and world models, meta-learning has been explored in three directions: test-time adaptation and scaling (Liu et al., 2025; Karageorgiou et al., 2026), generalization across scenarios via meta-regularized latent decomposition (Xiong et al., 2026) and contextual world models (Röder et al., 2025), and fine-grained credit assignment (Zhang et al., 2026; Savani et al., 2026). However, these methods either adapt only one side of the model while keeping the decoder fixed, or operate within a fixed reward space. In contrast, our RCM module jointly adapts the decoder and the world model through a bi-level formulation: the inner loop calibrates the decoder to the reachable space, and the outer loop updates the world model to generate decodable frames. This mutual adaptation eliminates target-set mismatch at its source, which static or one-sided adaptation cannot achieve.

3 Method

3.1 Problem Formulation

Given a dataset of continuous motion trajectories, we sample $H+1$ frames of RGB observations $X=\{x_t\}_{t=0}^{H}$ and the corresponding action sequence $U=\{u_t\}_{t=1}^{H}$, where x_t and u_t denote the observation and action at time step t. Taking (x_0, U) as the initial condition, the autoregressive world model π_θ generates the subsequent H frames by repeatedly predicting the next frame. At step t, for each candidate $i \in \{1, \dots, K\}$, the model conditions on its previously generated frame $\hat{x}_{i,t-1}$ (with $\hat{x}_{i,0} = x_0$) and the current action u_t to sample a token block:

$$z_{i,t} \sim \pi_\theta(\cdot \mid E(\hat{x}_{i,t-1}), Q(u_t)), \quad i \in \{1, \dots, K\}. \tag{1}$$

Each token block $z_{i,t}$ is decoded into a candidate next frame $\hat{x}_{i,t} = D(z_{i,t})$, and consists of L tokens $\{z_{i,t,l}\}_{l=1}^{L}$. The encoder E and quantizer Q are pre-trained and kept frozen during post-training. The

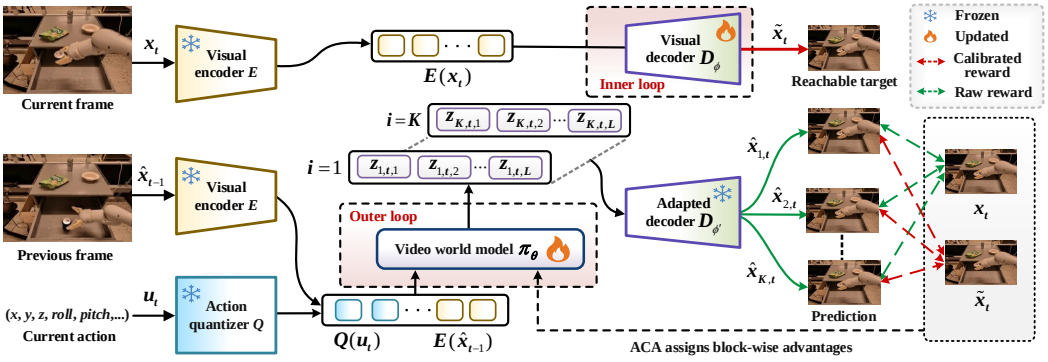


Figure 1: Overview of MetaReach. The inner loop adapts the decoder D_ϕ, while the outer loop updates the world model π_θ under $D_{\phi'}$. RCM calibrates rewards via reachable reconstructions, and ACA assigns block-wise advantages before gradient projection updates π_θ.

generated frame $\hat{x}_{i,t}$ serves as the condition for the next step, so each candidate produces its own trajectory autoregressively. Let $\mathcal{V} = \{1, \ldots, H\}$ denote the set of supervised time steps.

In the standard GRPO-based post-training paradigm (Shao et al., 2024; Wu et al., 2025; Zhu et al., 2026), only the world model π_θ is updated. A per-frame reward measures prediction accuracy:

$$r_{i,t}^{\text{raw}} = -d(\hat{x}_{i,t}, x_t), \tag{2}$$

where d combines pixel-level mean squared error and feature-level perceptual loss (Zhang et al., 2018). For each candidate i, a sequence-level reward $R_i = \frac{1}{|\mathcal{V}|}\sum_{t\in\mathcal{V}} r_{i,t}^{\text{raw}}$ is computed, and a group-relative advantage

$$A_i = \frac{R_i - \mu_R}{\sigma_R + \epsilon} \tag{3}$$

is obtained, where μ_R and σ_R are the mean and standard deviation of $\{R_i\}_{i=1}^K$, and ϵ is a small constant for numerical stability. The same advantage A_i is then assigned to all token blocks of candidate i.

This paradigm has two limitations: (1) *target-set mismatch*, where rewards against ground-truth frames are biased because the decoder cannot reproduce all details of ground-truth frames, distorting ranking; and (2) *long-horizon credit misallocation*, where a single sequence-level advantage is broadcast to all token blocks, ignoring the causal structure of autoregressive generation. To address both, we propose MetaReach, a bi-level meta-learning framework (Fig. 1) that uses an inner loop to adapt the decoder D to the world model's reachable space and an outer loop to update π_θ under the adapted decoder. This design, termed *Reachability-Calibrated Meta-learning (RCM)*, eliminates target-set mismatch at its source. We further introduce *Autoregressive Credit Assignment (ACA)*, which replaces the uniform advantage with a position-specific discounted return per token block, aligning credit with autoregressive dependencies. RCM corrects what the reward compares against, while ACA corrects how it is distributed over time.

3.2 Reachability-Calibrated Meta-learning (RCM)

Since rewards are computed on decoder outputs, an uncalibrated decoder introduces a fixed residual that corrupts candidate ranking. The decoder must therefore be calibrated before it can serve as a reliable evaluator. This motivates an inner-then-outer ordering: the inner loop first adapts the decoder to reconstruct frames within the world model's output distribution, and the outer loop then updates the world model under this calibrated decoder. Reversing the order would optimize the world model against a biased evaluator, which cannot eliminate the mismatch.

Inner loop: decoder adaptation. Let ϕ denote the trainable parameters of the decoder D. At the beginning of each training iteration, ϕ is inherited from the previous iteration. The inner loop updates ϕ so that the decoder reconstructs the current frame x_t from its encoded representation:

$$\mathcal{L}_{\text{inner}}(\phi) = \mathbb{E}_{x_t}\left[\|D_\phi(E(x_t)) - x_t\|_2^2\right]. \tag{4}$$

Algorithm 1 Training Process of MetaReach

Require: Frozen encoder E and quantizer Q; trainable decoder D_ϕ and world model π_θ.
1: Initialize ϕ and θ from pre-trained checkpoints.
2: **for** each trajectory (X, U) in a batch **do**
3: **// Inner loop:** update ϕ for N_{inner} steps via Eq. (4) to obtain $D_{\phi'}$;
4: Compute reachable reconstruction $\tilde{x}_t = D_{\phi'}(E(x_t))$ via Eq. (5);
5: **// Outer loop:** for each step t, sample K candidate token blocks $z_{i,t}$ from $\pi_\theta(\cdot \mid E(\hat{x}_{i,t-1}), Q(u_t))$ via Eq. (1), yielding $\hat{x}_{i,t} = D_{\phi'}(z_{i,t})$;
6: Compute raw reward $r^{\text{raw}}_{i,t}$ via Eq. (2) and calibrated reward $r^{\text{RC}}_{i,t}$ via Eq. (6);
7: Compute advantages $A^{\text{raw}}_{i,t}$ and $A^{\text{RC}}_{i,t}$ via Eq. (10);
8: Compute g_0 and g_1 via Eq. (11);
9: Compute projected gradient g^* via Eq. (13);
10: Update $\theta \leftarrow \theta - \eta g^*$;
11: **end for**
12: **return** Optimized π_θ and the final adapted decoder $D_{\phi'}$.

The action u_t is not used here because reconstructing x_t depends only on its encoded representation. We take N_{inner} gradient steps with a small learning rate to avoid overfitting to a single trajectory, obtaining an adapted decoder $D_{\phi'}$. This adaptation is performed for every trajectory in the batch, so ϕ evolves across iterations, while the outer loop treats ϕ' as fixed.

Outer loop: world model update. With the adapted decoder $D_{\phi'}$, the world model samples K candidate token blocks $z_{i,t}$ following the same conditional distribution as in Eq. (1), and decodes them with $D_{\phi'}$ to obtain predictions $\hat{x}_{i,t} = D_{\phi'}(z_{i,t})$. The reachable reconstruction of the ground-truth frame x_t is

$$\tilde{x}_t = D_{\phi'}(E(x_t)), \tag{5}$$

which lies inside the decoder's reachable space. We define the calibrated reward as

$$r^{\text{RC}}_{i,t} = -d(\hat{x}_{i,t}, \tilde{x}_t), \tag{6}$$

where d is the same distance function as in Eq. (2). For each candidate i, a sequence-level reward is computed by averaging $r^{\text{RC}}_{i,t}$ over $\mathcal{V}$, and a group-relative advantage A^{RC}_i is obtained by normalizing within the group, analogous to Eq. (3). The outer-loop loss with the calibrated reward is

$$\mathcal{L}^{\text{RC}}_{\text{outer}}(\theta) = -\frac{1}{K}\sum_{i=1}^{K} A^{\text{RC}}_i \sum_{t=1}^{H} \log \pi_\theta(z_{i,t} \mid E(\hat{x}_{i,t-1}), Q(u_t)), \tag{7}$$

where $\log \pi_\theta(z_{i,t} \mid \cdot)$ denotes the sum of log probabilities over all tokens in block $z_{i,t}$. Analogously, the raw reward $r^{\text{raw}}_{i,t}$ yields $\mathcal{L}^{\text{raw}}_{\text{outer}}(\theta)$ with advantage A^{raw}_i. Only the world model parameters θ are updated in this stage, while ϕ' remains frozen, consistent with the standard bi-level optimization where the lower-level solution is treated as fixed during the upper-level update. The sequence-level formulation in Eq. (7) still broadcasts the same advantage to all token blocks; the ACA strategy below refines it by assigning a position-specific advantage to each block.

3.3 Autoregressive Credit Assignment (ACA)

The outer-loop loss in Eq. (7) broadcasts a sequence-level advantage to all token blocks. ACA instead assigns a block-wise advantage $A^{\text{RC}}_{i,t}$ to each block, yielding the refined loss

$$\mathcal{L}^{\text{RC}}_{\text{outer}}(\theta) = -\frac{1}{K}\sum_{i=1}^{K}\sum_{t=1}^{H} A^{\text{RC}}_{i,t}\, \ell_{i,t}, \tag{8}$$

where

$$\ell_{i,t} = \sum_{l=1}^{L} \log \pi_\theta(z_{i,t,l} \mid E(\hat{x}_{i,t-1}), Q(u_t), z_{i,t,<l}) \tag{9}$$

is the log-probability of block $z_{i,t}$, conditioned on the encoded previous frame and the preceding tokens within the same block.

The block-wise advantage accumulates a discounted return and normalizes across candidates at the same time step:

$$Z_{i,t}^{\mathrm{RC}} = \sum_{t'=t}^{H} \gamma^{t'-t} r_{i,t'}^{\mathrm{RC}}, \quad A_{i,t}^{\mathrm{RC}} = \frac{Z_{i,t}^{\mathrm{RC}} - \mu_t^{\mathrm{RC}}}{\sigma_t^{\mathrm{RC}} + \epsilon}, \tag{10}$$

where γ is the discount factor, and $\mu_t^{\mathrm{RC}}, \sigma_t^{\mathrm{RC}}$ are the mean and standard deviation of $\{Z_{i,t}^{\mathrm{RC}}\}_{i=1}^{K}$ across the K candidates at time step t. Replacing $r_{i,t}^{\mathrm{RC}}$ with $r_{i,t}^{\mathrm{raw}}$ in Eq. (10) yields the raw variant $A_{i,t}^{\mathrm{raw}}$ and the corresponding loss $\mathcal{L}_{\mathrm{outer}}^{\mathrm{raw}}$. Their gradients with respect to θ are denoted as

$$g_1 = \nabla_\theta \mathcal{L}_{\mathrm{outer}}^{\mathrm{RC}}, \qquad g_0 = \nabla_\theta \mathcal{L}_{\mathrm{outer}}^{\mathrm{raw}}. \tag{11}$$

Gradient projection. The calibrated reward may shift the update direction away from the original objective. To preserve progress on the original objective, we seek a direction g^* closest to g_1 that still achieves first-order descent on the original objective at least as large as g_0:

$$g^* = \arg\min_{g} \frac{1}{2}\|g - g_1\|^2 \quad \text{s.t.} \quad \langle g, g_0 \rangle \geq \|g_0\|^2. \tag{12}$$

This constraint prevents the calibrated signal from degrading the original training progress. The problem admits a closed-form solution:

$$g^* = g_1 + \lambda^* g_0, \quad \lambda^* = \frac{[\|g_0\|_2^2 - \langle g_1, g_0 \rangle]_+}{\|g_0\|_2^2}, \tag{13}$$

where $[x]_+ = \max(x, 0)$. The world model parameters are then updated as

$$\theta \leftarrow \theta - \eta g^*, \tag{14}$$

where η is the learning rate. When g_1 already satisfies the constraint, $\lambda^* = 0$ and the calibrated direction is used directly; otherwise, only the minimal multiple of g_0 needed to restore the constraint is added. If $\|g_0\|_2 = 0$, the constraint is trivial and $g^* = g_1$.

Alternating optimization. The inner and outer loops alternate iteratively during the post-training phase. Each training iteration adapts the decoder ϕ for N_{inner} steps and then updates the world model θ exactly once using the projected gradient g^*. This bi-level meta-learning procedure jointly calibrates the decoder to the reachable space and aligns the world model with the calibrated evaluator effectively. All hyperparameters N_{inner}, γ, and ϵ are provided in the experimental setup.

This design clearly distinguishes ACA from prior credit assignment methods in two key aspects. First, unlike step-level methods designed for diffusion or flow-matching trajectories (Savani et al., 2026; Ding & Ye, 2026), ACA operates at the fine-grained token-block level in autoregressive video generation, where each block conditions all subsequent blocks and thus naturally corresponds to a distinct position-specific future horizon. Second, unlike methods relying on auxiliary value networks, learned reward estimators, or tree-structured rollouts (Savani et al., 2026), ACA derives credit directly from the same discounted return that ultimately drives the policy update. These properties make ACA fully compatible with GRPO and complementary to RCM: while RCM corrects what the reward compares against, ACA corrects how the reward is distributed over time.

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate our method on three action-conditioned video prediction benchmarks. (1) RT-1 (Brohan et al., 2023) contains 87,212 real-robot manipulation trajectories at 256×320 resolution with 13-dimensional actions; we use 86,340 for training and 872 for evaluation. (2) RoboNet (Dasari et al., 2019) comprises 162,417 multi-platform trajectories; we retain the official 256 test trajectories and use the remaining 162,161 for training, with images resized to 64×64 and 5-dimensional actions. (3) VP2-RoboSuite (Tian et al., 2023) provides 5,000 simulated manipulation demonstrations, split into 4,478 training and 522 validation trajectories, with 64×64 images and 4-dimensional actions. For all datasets, one fixed clip per held-out trajectory is used to ensure identical test inputs across methods.

Dataset	Method	Mean over the 15-Frame Prediction Horizon				Final LPIPS↓
		LPIPS↓	MSE↓	PSNR↑	SSIM↑	
RT-1	Baseline (Wu et al., 2025)	0.23645	0.01987	17.792	0.70412	0.26783
	RLVR-World (Wu et al., 2025)	0.22591	0.01824	18.083	0.72056	0.24987
	Seq-GRPO (Shao et al., 2024)	0.22014	0.01731	18.372	0.72394	0.24652
	WorldCompass (Wang et al., 2026b)	0.21506	0.01661	18.612	0.73087	0.23914
	Self Forcing (Huang et al., 2025)	0.21583	0.01672	18.548	0.72945	0.24036
	MetaReach (w/o RCM)	0.21487	0.01663	18.591	0.72987	0.23872
	MetaReach (w/o ACA)	0.21241	0.01628	18.734	0.73312	0.23418
	MetaReach	**0.20856**	**0.01582**	**18.964**	**0.73901**	**0.22631**
RoboNet	Baseline (Wu et al., 2024)	0.08043	0.006058	23.942	0.83401	0.09587
	Seq-GRPO (Shao et al., 2024)	0.07987	0.005991	24.008	0.83488	0.09471
	WorldCompass (Wang et al., 2026b)	0.07942	0.005944	24.121	0.83641	0.09388
	MetaReach (w/o RCM)	0.07958	0.005961	24.058	0.83547	0.09421
	MetaReach (w/o ACA)	0.07971	0.005973	24.032	0.83558	0.09435
	MetaReach	**0.07901**	**0.005902**	**24.214**	**0.83756**	**0.09294**

Table 1: Quantitative comparison of video prediction quality over a 15-frame prediction horizon. Bold indicates the best mean across five paired seeds per dataset.

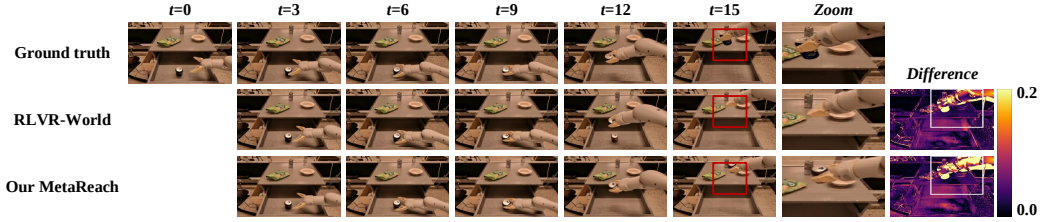


Figure 2: Qualitative comparison on RT-1. Left: frames from $t = 0$ to $t = 15$ with a stride of 3. Right: final-frame zoom and difference map. RLVR-World (Wu et al., 2025) fails to grasp the can and loses it, while MetaReach grasps and lifts it with smaller deviations.

Implementation Details. The world model π_θ follows the same autoregressive Transformer architecture as RLVR-World (Wu et al., 2025). We initialize it with the RLVR-World visual Transformer for RT-1 (Wu et al., 2025), and with action-conditioned iVideoGPT checkpoints for RoboNet and VP2-RoboSuite (Wu et al., 2024). MetaReach runs 200 iterations, each comprising N_{inner}=3 inner-loop steps (AdamW, lr 1×10^{-4}) that adapt the decoder, followed by one outer-loop update of π_θ (lr 5×10^{-5}, weight decay 0.01, grad clipping 1.0). The decoder parameters are carried across iterations. We set K=16, γ=0.95, and ϵ=1×10^{-8}. The total batch size is 128, implemented via gradient accumulation over 8 micro-batches of 16. RT-1 uses 80 visual tokens per frame (sequence length 2,024); RoboNet and VP2-RoboSuite use 16 tokens per frame (sequence lengths 1,024 and 649).

Evaluation Metrics. For RT-1 and RoboNet, models predict the next 15 frames given an initial frame and future actions. VP2-RoboSuite uses the first two frames to predict the subsequent 8 frames. We evaluate full-rollout fidelity using LPIPS (Zhang et al., 2018), MSE, PSNR (Horé & Ziou, 2010), and SSIM (Wang et al., 2004), and separately report LPIPS on the final predicted frame. To evaluate reward quality, we adopt pairwise accuracy (Christiano et al., 2017; Lambert et al., 2024) and top-1 accuracy (Cobbe et al., 2021; Lightman et al., 2024), which directly characterize the supervision consumed by group-relative optimization. All experiments run on 4× 80G A100 GPUs, with results averaged over the last five checkpoints.

4.2 Main Results

Quantitative Comparison. On RT-1, MetaReach reduces LPIPS by 7.7% and MSE by 13.3% over RLVR-World (Wu et al., 2025), also surpassing Seq-GRPO (Shao et al., 2024), WorldCompass (Wang et al., 2026b) and Self Forcing (Huang et al., 2025). These gains stem from the fact that

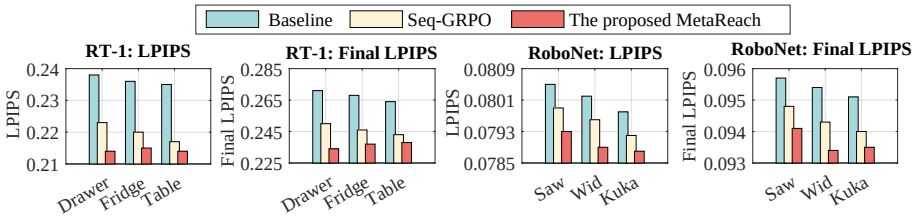


Figure 3: Generalization to held-out scenes over a 15-frame prediction horizon. MetaReach outperforms Baseline and Seq-GRPO (Shao et al., 2024) in both LPIPS and final-frame LPIPS.

Meta-Learning	N_{inner}	LPIPS↓	MSE↓	PSNR↑	SSIM↑	Final LPIPS↓
×	0	0.21750	0.01705	18.412	0.72750	0.24120
✓	1	0.21020	0.01610	18.850	0.73650	0.22950
✓	3	**0.20856**	**0.01582**	**18.964**	**0.73901**	**0.22631**
✓	5	0.20840	0.01579	18.975	0.73920	0.22600
✓	10	0.20895	0.01588	18.940	0.73850	0.22680

Table 2: Effect of inner-loop steps N_{inner} in RCM on RT-1. N_{inner}=0 disables decoder adaptation (RCM without inner loop). Performance improves up to 3 steps and saturates thereafter.

prior methods freeze the decoder and compute rewards against ground-truth frames, introducing a fixed residual that distorts ranking (target-set mismatch), and broadcast a sequence-level advantage uniformly to all token blocks, ignoring autoregressive causality (long-horizon credit misallocation). MetaReach instead calibrates the decoder to the reachable space via the inner loop and assigns block-wise advantages via ACA. Ablations confirm that removing RCM degrades LPIPS by 3.0% and final-frame LPIPS by 5.5%, while removing ACA degrades them by 1.9% and 3.5%, showing that RCM improves ranking fidelity and ACA enhances temporal consistency.

Qualitative Evaluation. Fig. 2 compares RT-1 rollouts over 15 frames among ground truth, RLVR-World (Wu et al., 2025), and MetaReach. The ground truth shows the robot grasping and lifting the can. RLVR-World fails to grasp it, and in the final frame the can is neither grasped nor visible, indicating inconsistent object representation and incorrect interaction logic. In contrast, MetaReach successfully grasps and lifts the can, closely matching the ground truth with smaller deviations. This indicates that the improvement lies not only in pixel-level fidelity but in preserving key interaction semantics, which is critical for downstream control.

Generalization to Held-Out Scenes. We evaluate generalization to unseen scenes over a 15-frame prediction horizon. Fig. 3 compares MetaReach with the Baseline and Seq-GRPO (Shao et al., 2024) on three held-out RT-1 scenes (Drawer, Fridge, Table) and three RoboNet scenes (Sawyer, WidowX, Kuka). Across all six scenes, MetaReach achieves the lowest LPIPS and final-frame LPIPS. On RT-1 Drawer retrieval, final-frame LPIPS drops from 0.250 (Seq-GRPO) to 0.234 and average LPIPS from 0.223 to 0.214, with consistent gains on RoboNet. The advantage is more pronounced on the final frame than on the sequence average, indicating that ACA's block-wise credit assignment mitigates error accumulation over long horizons. These results confirm that MetaReach generalizes robustly to unseen scenes and validate the complementary roles of RCM and ACA.

4.3 Ablation Study

Effect of Inner-Loop Adaptation in RCM. Table 2 varies the number of inner-loop steps N_{inner} in the RCM module on RT-1 over a 15-frame horizon. Setting N_{inner}=0 disables decoder adaptation, leaving the decoder fixed and reducing RCM to a static calibration without meta-learning. Performance improves substantially from 0 to 3 steps, with final-frame LPIPS dropping from 0.24120 to 0.22631, confirming that adapting the decoder to the world model's output distribution is essential for reliable reward comparison. Increasing N_{inner} beyond 3 yields no further gain, and 10 steps slightly degrades performance, indicating that a small number of adaptation steps is sufficient and that excessive steps may overfit a single trajectory. We therefore adopt N_{inner}=3 by default.

Method	LPIPS t=1	LPIPS t=5	LPIPS t=10	LPIPS t=15	PSNR t=15	SSIM t=15
w/o ACA	0.082	0.128	0.195	0.245	17.5	0.69
w/ ACA	0.080	0.118	0.178	0.226	18.4	0.72
Gain (%)	2.4	7.8	8.7	7.8	+0.9	+0.03

Table 3: Per-frame prediction quality on RT-1 over a 15-frame horizon, comparing MetaReach with and without ACA. The gain grows with t and peaks at the final frame, confirming that ACA mitigates error accumulation in long-horizon autoregressive generation.

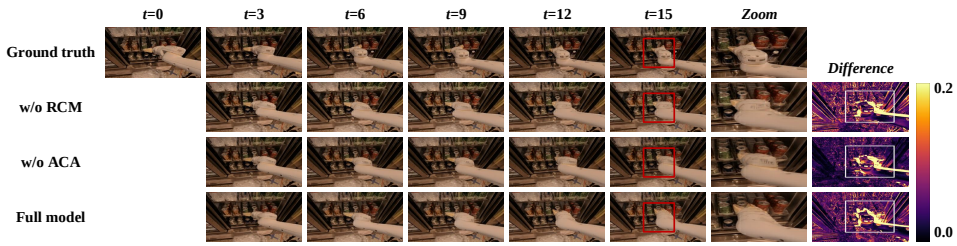


Figure 4: Qualitative ablation on RT-1. Without RCM, the wrist grid and body logo disappear and the end-effector becomes blurry; without ACA, motion stalls mid-sequence and never contacts the target. The full model best matches the ground truth.

Effect of ACA on Per-Frame Quality. Table 3 compares MetaReach with and without ACA on RT-1 over a 15-frame horizon. On early frames the two variants perform similarly, but the gap widens steadily with t: LPIPS at t=10 improves from 0.195 to 0.178, and at the final frame from 0.245 to 0.226, with PSNR improving by 0.9 dB and SSIM by 0.03. This widening trend confirms that ACA mitigates error accumulation in autoregressive generation. By assigning credit to each token block according to its future horizon, ACA ensures that early predictions are optimized more carefully, which benefits all subsequent frames and yields the largest gains at the end of the rollout. These results provide evidence for ACA's role in addressing long-horizon credit misallocation.

Qualitative Comparison of Ablation Variants. Fig. 4 further presents an ablation scene in which all three configurations keep the robotic arm in front of the shelf without directional errors, making the overall visual difference smaller than in the previous scene. The differences lie in two aspects. First, appearance restoration: in the w/o RCM variant, the grid structure on the wrist and the body logo disappear, and the end-effector degenerates into a smooth, blurry shape. Second, motion progression: the w/o ACA variant exhibits several nearly static frames in the middle of the sequence and, although motion resumes afterward, never establishes contact with the target object. The full model matches the ground-truth motion rhythm most closely.

5 Conclusion

We propose MetaReach, a bi-level meta-learning framework for post-training video world models, where RCM calibrates the decoder to the reachable space through inner-outer loops and ACA assigns block-wise advantages to align credit with autoregressive dependencies. Experiments on RT-1, RoboNet, and VP2-RoboSuite over a 15-frame horizon demonstrate consistent gains over GRPO-based baselines. Although MetaReach achieves these improvements, two limitations remain. First, the reachable target is defined by a frozen tokenizer and decoder; when the visual representation is continuous or the decoder is trained jointly, the reachable space is no longer fixed and the calibration mechanism must be redefined. Second, the reward relies on per-frame appearance distance, which favors predictions that under-represent motion and may reward static or blurry outputs. RCM corrects which candidate wins within each group, but it does not change what the metric rewards. Future work includes extending the reachable set to diffusion-based decoders, exploring continuous latent calibration, incorporating motion-aware verifiers such as optical flow or trajectory consistency into the calibrated reward, generalizing RCM and ACA to other modalities and longer horizons, and integrating them with alternative policy optimization algorithms beyond GRPO.

REFERENCES

Isabela Albuquerque, Jessica Schrouff, David Warde-Farley, Ali Taylan Cemgil, Sven Gowal, and Olivia Wiles. Evaluating model bias requires characterizing its mistakes. In *Forty-first International Conference on Machine Learning*, Proceedings of the 41st International Conference on Machine Learning, pp. 938–954, 2024.

Antreas Antoniou, Harrison Edwards, and Amos Storkey. How to train your MAML. In *International Conference on Learning Representations*, 2019.

Hongzhe Bi, Hengkai Tan, Shenghao Xie, Zeyuan Wang, Shuhe Huang, Haitian Liu, Ruowen Zhao, Yao Feng, Chendong Xiang, Yinze Rong, Hongyan Zhao, Hanyu Liu, Zhizhong Su, Lei Ma, Hang Su, and Jun Zhu. Motus: A unified latent action world model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, et al. RT-1: Robotics transformer for real-world control at scale. In *Robotics: Science and Systems XIX*, 2023.

Maxime Burchi and Radu Timofte. Accurate and efficient world modeling with masked latent transformers. In *Proceedings of the 42nd International Conference on Machine Learning*, 2025.

Paul F. Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems*, pp. 4299–4307, 2017.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Lior Cohen, Kaixin Wang, Bingyi Kang, and Shie Mannor. Improving token-based world models with parallel observation prediction. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 9138–9160, 2024.

Chenxu Dang, Haiyan Liu, Jason Bao, Pei An, Xinyue Tang, An Pan, Jie Ma, Bingchuan Sun, and Yan Wang. SparseWorld: A flexible, adaptive, and efficient 4d occupancy world model powered by sparse and dynamic queries. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence*, pp. 3497–3505, 2026.

Sudeep Dasari, Frederik Ebert, Stephen Tian, Suraj Nair, Bernadette Bucher, Karl Schmeckpeper, Siddharth Singh, Sergey Levine, and Chelsea Finn. RoboNet: Large-scale multi-robot learning. In *Proceedings of the 3rd Annual Conference on Robot Learning*, volume 100, pp. 885–897, 2019.

Jingtao Ding, Yunke Zhang, Yu Shang, Yuheng Zhang, Zefang Zong, Jie Feng, Yuan Yuan, Hongyuan Su, Nian Li, Nicholas Sukiennik, Fengli Xu, and Yong Li. Understanding world or predicting future? A comprehensive survey of world models. *ACM Comput. Surv.*, 58(3): 57:1–57:38, 2026.

Zheng Ding and Weirui Ye. TreeGRPO: Tree-advantage GRPO for online RL post-training of diffusion models. In *Proceedings of the Fourteenth International Conference on Learning Representations*, 2026.

Tuo Feng, Wenguan Wang, and Yi Yang. A survey of world models for autonomous driving. *arXiv preprint arXiv:2501.11260*, 2025.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International Conference on Machine Learning*, 2017.

Mitchell Goff, Greg Hogan, George Hotz, Armand du Parc Locmaria, Kacper Raczy, Harald Schäfer, Adeeb Shihadeh, Weixing Zhang, and Yassine Yousfi. Learning to drive from a world model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 1964–1973, 2025.

Alain Horé and Djemel Ziou. Image quality metrics: PSNR vs. SSIM. In *20th International Conference on Pattern Recognition*, pp. 2366–2369, 2010.

Siqiao Huang, Jialong Wu, Qixing Zhou, Shangchen Miao, and Mingsheng Long. Vid2World: Crafting video diffusion models to interactive world models. In *Proceedings of the 14th International Conference on Learning Representations*, 2026.

Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self forcing: Bridging the train-test gap in autoregressive video diffusion. In *Advances in Neural Information Processing Systems*, 2025.

Miles Hutson, Isaac Kauvar, and Nick Haber. Policy-shaped prediction: Avoiding distractions in model-based reinforcement learning. In *Advances in Neural Information Processing Systems*, 2024.

Feiyang Jia, Lin Liu, Ziying Song, Caiyan Jia, Hangjun Ye, Xiaoshuai Hao, and Long Chen. DriveWorld-VLA: Unified latent-space world modeling with vision–language–action for autonomous driving. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026.

Dimitrios Karageorgiou, Symeon Papadopoulos, Ioannis Kompatsiaris, and Efstratios Gavves. TANGO: Test-time noise guided adaptation for realistic autoregressive video generation. In *Proceedings of the European Conference on Computer Vision*, 2026.

Dongwon Kim, Gawon Seo, Jinsung Lee, Minsu Cho, and Suha Kwak. Planning in 8 tokens: A compact discrete tokenizer for latent world model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

Nathan Lambert, Valentina Pyatkin, Jacob Morrison, Lester James V. Miranda, Bill Yuchen Lin, Khyathi Chandu, Nouha Dziri, Sachin Kumar, Tom Zick, Yejin Choi, Noah A. Smith, and Hannaneh Hajishirzi. Rewardbench: Evaluating reward models for language modeling. *arXiv preprint arXiv:2403.13787*, 2024.

Sikuang Li, Chen Yang, Jiemin Fang, Taoran Yi, Jia Lu, Jiazhong Cen, Lingxi Xie, Wei Shen, and Qi Tian. Worldgrow: Generating infinite 3d world. In *Proceedings of the 40th AAAI Conference on Artificial Intelligence*, pp. 6433–6441, 2026.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *International Conference on Learning Representations*, 2024.

Fangfu Liu, Hanyang Wang, Yimo Cai, Kaiyan Zhang, Xiaohang Zhan, and Yueqi Duan. Video-T1: Test-time scaling for video generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 18671–18681, 2025.

Vincent Micheli, Eloi Alonso, and François Fleuret. Efficient world models with context-aware tokenization. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 35623–35638, 2024.

Yosuke Nishimoto and Takashi Matsubara. Object-centric world models for causality-aware reinforcement learning. In Sven Koenig, Chad Jenkins, and Matthew E. Taylor (eds.), *Proceedings of the Fortieth AAAI Conference on Artificial Intelligence*, pp. 24585–24593, 2026.

Ryan Po, Yotam Nitzan, Richard Zhang, Berlin Chen, Tri Dao, Eli Shechtman, Gordon Wetzstein, and Xun Huang. Long-context state-space video world models. In *Proceedings of IEEE/CVF International Conference on Computer Vision*, pp. 8733–8744, 2025.

Aditya A. Ramesh, Kenny John Young, Louis Kirsch, and Jürgen Schmidhuber. Sequence compression speeds up credit assignment in reinforcement learning. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 42104–42126, 2024.

Zhongwei Ren, Yunchao Wei, Xun Guo, Yao Zhao, Bingyi Kang, Jiashi Feng, and Xiaojie Jin. Videoworld: Exploring knowledge learning from unlabeled videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 29029–29039, 2025.

Marc Rigter, Minqi Jiang, and Ingmar Posner. Reward-free curricula for training robust world models. In *Proceedings of the 12th International Conference on Learning Representations*, 2024.

Frank Röder, Jan Benad, Manfred Eppe, and Pradeep Kr. Banerjee. Dynamics-aligned latent imagination in contextual world models for zero-shot generalization. In *Advances in Neural Information Processing Systems*, 2025.

Mohammad Reza Samsami, Artem Zholus, Janarthanan Rajendran, and Sarath Chandar. Mastering memory tasks with world models. In *Proceedings of the Twelfth International Conference on Learning Representations*, 2024.

Yash Savani, Branislav Kveton, Yuchen Liu, Yilin Wang, Jing Shi, Subhojyoti Mukherjee, Nikos Vlassis, and Krishna Kumar Singh. Stepwise credit assignment for GRPO on flow-matching models. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 42007–42017, 2026.

Nedko Savov, Naser Kazemi, Deheng Zhang, Danda Pani Paudel, Xi Wang, and Luc Van Gool. StateSpaceDiffuser: Bringing long context to diffusion world models. In *Advances in Neural Information Processing Systems*, 2025.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Stephen Tian, Chelsea Finn, and Jiajun Wu. A control-centric benchmark for video prediction. In *Proceedings of the 11th International Conference on Learning Representations*, 2023.

Keyon Vafa, Justin Y. Chen, Ashesh Rambachan, Jon M. Kleinberg, and Sendhil Mullainathan. Evaluating the world model implicit in a generative model. In *Advances in Neural Information Processing Systems*, 2024.

Mudit Verma and Katherine Metcalf. Hindsight PRIORs for reward learning from human preferences. In *Proceedings of the 21th International Conference on Learning Representations*, 2024.

Hanyu Wang, Saksham Suri, Yixuan Ren, Hao Chen, and Abhinav Shrivastava. LARP: Tokenizing videos with a learned autoregressive generative prior. In *Proceedings of the Thirteenth International Conference on Learning Representations*, 2025a.

Hongwei Wang, Yangru Huang, Guangyao Chen, Xu Wang, and Yi Jin. Beyond single-speed reasoning: Coordinating fast and slow dynamics for efficient world modeling. In *Proceedings of the Fortieth AAAI Conference on Artificial Intelligence*, pp. 26215–26223, 2026a.

Qi Wang, Zhipeng Zhang, Baao Xie, Xin Jin, Yunbo Wang, Shiyu Wang, Liaomo Zheng, Xiaokang Yang, and Wenjun Zeng. Disentangled world models: Learning to transfer semantic knowledge from distracting videos for reinforcement learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2599–2608, 2025b.

Yucen Wang, Rui Yu, Shenghua Wan, Le Gan, and De-Chuan Zhan. FOUNDER: Grounding foundation models in world models for open-ended embodied decision making. In *Proceedings of the 42nd International Conference on Machine Learning*, 2025c.

Yuqing Wang, Zhijie Lin, Yao Teng, Yuanzhi Zhu, Shuhuai Ren, Jiashi Feng, and Xihui Liu. Bridging continuous and discrete tokens for autoregressive visual generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 18596–18605, 2025d.

Zehan Wang, Tengfei Wang, Haiyu Zhang, Xuhui Zuo, Junta Wu, Haoyuan Wang, Wenqiang Sun, Zhenwei Wang, Chenjie Cao, Hengshuang Zhao, Chunchao Guo, and Zhou Zhao. WorldCompass: Reinforcement learning for long-horizon world models. In *Proceedings of the 43rd International Conference on Machine Learning (ICML)*, 2026b.

Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4): 600–612, April 2004.

Jialong Wu, Shaofeng Yin, Ningya Feng, Xu He, Dong Li, Jianye Hao, and Mingsheng Long. iVideoGPT: Interactive VideoGPTs are scalable world models. In *Advances in Neural Information Processing Systems*, 2024.

Jialong Wu, Shaofeng Yin, Ningya Feng, and Mingsheng Long. RLVR-World: Training world models with reinforcement learning. In *Advances in Neural Information Processing Systems*, 2025.

Tianze Xia, Yongkang Li, Lijun Zhou, Jingfeng Yao, Kaixin Xiong, Haiyang Sun, Bing Wang, Kun Ma, Guang Chen, Hangjun Ye, Wenyu Liu, and Xinggang Wang. DriveLaW: Unifying planning and video generation in a latent driving world. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 39701–39712, 2026.

Xuantang Xiong, Ni Mu, Runpeng Xie, Senhao Yang, Yaqing Wang, Lexiang Wang, Yao Luan, Siyuan Li, Shuang Xu, Yiqin Yang, and Bo Xu. MrCoM: A meta-regularized world-model generalizing across multi-scenarios. In *Proceedings of the Fortieth AAAI Conference on Artificial Intelligence*, pp. 27180–27187, 2026.

Wilson Yan, Volodymyr Mnih, Aleksandra Faust, Matei Zaharia, Pieter Abbeel, and Hao Liu. ElasticTok: Adaptive tokenization for image and video. In *Proceedings of the Thirteenth International Conference on Learning Representations*, 2025.

Fuxiang Yang, Donglin Di, Lulu Tang, Xuancheng Zhang, Lei Fan, Hao Li, Chen Wei, Tonghua Su, and Baorui Ma. Chain of world: World model thinking in latent motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026a.

Yuxue Yang, Lue Fan, Ziqi Shi, Junran Peng, Feng Wang, and Zhaoxiang Zhang. NeoVerse: Enhancing 4d world model with in-the-wild monocular videos. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 40340–40351, 2026b.

Wei Yao, Haian Yin, Shangzhi Zeng, and Jin Zhang. Overcoming lower-level constraints in bilevel optimization: A novel approach with regularized gap functions. In *Proceedings of the Thirteenth International Conference on Learning Representations*, 2025.

Minjong Yoo, Jinwoo Jang, Sihyung Yoon, and Honguk Woo. World model implanting for test-time adaptation of embodied agents. In *Proceedings of the Forty-second International Conference on Machine Learning*, 2025.

Kaiwen Zhang, Zhenyu Tang, Xiaotao Hu, Xingang Pan, Xiaoyang Guo, Yuan Liu, Jingwei Huang, Li Yuan, Qian Zhang, Xiao-Xiao Long, Xun Cao, and Wei Yin. Epona: Autoregressive diffusion world model for autonomous driving. In *Proceedings of IEEE/CVF International Conference on Computer Vision*, pp. 27220–27230, 2025.

Mu Zhang, Tianren Ma, Yunfan Liu, Kun Hu, and Qixiang Ye. RebRL: Reinforcing discrete visual diffusion models with rebalanced timestep credits. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 43135–43144, 2026.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 586–595, 2018.

Weipu Zhang, Gang Wang, Jian Sun, Yetian Yuan, and Gao Huang. STORM: Efficient stochastic transformer based world models for reinforcement learning. In *Advances in Neural Information Processing Systems*, 2023.

Guosheng Zhao, Xiaofeng Wang, Zheng Zhu, Xinze Chen, Guan Huang, Xiaoyi Bao, and Xingang Wang. DriveDreamer-2: LLM-enhanced world models for diverse driving video generation. In *Proceedings of the 39th AAAI Conference on Artificial Intelligence*, pp. 10412–10420, 2025.

Fangqi Zhu, Zhengyang Yan, Zicong Hong, Quanxin Shou, Xiao Ma, and Song Guo. WMPO: World model-based policy optimization for vision-language-action models. In *Proceedings of the 14th International Conference on Learning Representations*, 2026.

Luisa Zintgraf, Kyriacos Shiarlis, Vitaly Kurin, Katja Hofmann, and Shimon Whiteson. Fast context adaptation via meta-learning. In *International Conference on Machine Learning*, 2019.