Enhancing Vision-Language Model Training with Reinforcement Learning in Synthetic Worlds for Real-World Success

George Bredis
T-Tech
Moscow, Russia
georgy.bredis@gmail.com

Stanislav Dereka
T-Tech
Moscow, Russia
st.dereka@gmail.com

Viacheslav Sinii
T-Tech
Moscow, Russia
siniy.vyacheslav@gmail.com

Ruslan Rakhimov
T-Tech
Moscow, Russia
u.rakhimov@t-tech.dev

Daniil Gavrilov
T-Tech
Moscow, Russia
olavursky@gmail.com

ABSTRACT

Interactive multimodal agents must turn raw visual observations into reliable sequences of structured, language-conditioned actions, yet training such competence under long horizons and sparse feedback remains brittle. We present VL-DAC, a lightweight reinforcement learning recipe for vision-language agents that is hyperparameter-robust and easy to deploy. VL-DAC performs PPO updates at the token level for actions while learning a step-level value function. This decoupling removes unstable weighting terms and yields faster, more reliable convergence without introducing extra tuning knobs. Training a single VLM in one cheap synthetic environment at a time (MiniWorld, Gym-Cards, ALFWorld, or WebShop) produces policies that transfer beyond their training simulators: +50% relative on BALROG (agentic control), +5% relative on the hardest split of VSI-Bench (spatial planning), and +2% on VisualWebBench (web navigation), with no loss in general image understanding. Together, these results show that a simple, stable RL procedure can train vision–language agents entirely in simulation while delivering measurable gains on agentic, spatial-reasoning, and web-navigation benchmarks.

KEYWORDS

vision-language agents, reinforcement learning, GAAI

ACM Reference Format:
George Bredis, Stanislav Dereka, Viacheslav Sinii, Ruslan Rakhimov, and Daniil Gavrilov. 2026. Enhancing Vision-Language Model Training with Reinforcement Learning in Synthetic Worlds for Real-World Success. In *Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), Paphos, Cyprus, May 25 – 29, 2026*, IFAAMAS, 10 pages. https://doi.org/10.65109/

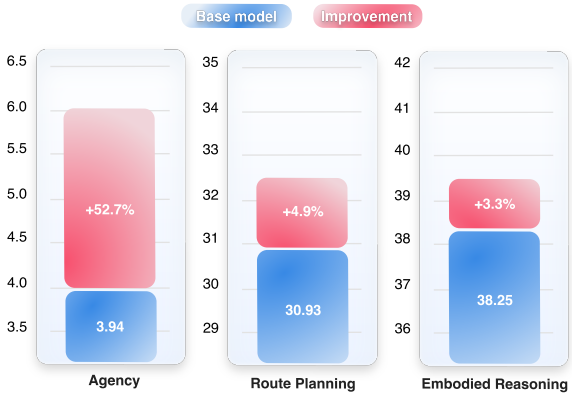

Figure 1: Skill transfer after synthetic training. Our method, VL-DAC, improves agentic control, spatial planning, and embodied reasoning on BALROG, VSI-Bench, and ERQA. It demonstrates effective transfer from synthetic environments to specific-skill benchmarks.

1 INTRODUCTION

Large language models (LLMs) behave like capable single-turn agents in text-only domains, where reinforcement learning (RL) can be applied without manual annotation [9, 18]. Yet these models still struggle when tasks require reasoning over many steps, exposing persistent limitations in long-horizon credit assignment and agentic control, arguably the main bottlenecks to more general capabilities. These challenges intensify for vision-language models (VLMs) ([29], [6]): in addition to planning across multiple steps, a VLM must parse a constantly changing visual stream. While state-of-the-art VLMs excel at describing static images and videos, they struggle to decide *what to do next* in interactive or dynamic scenes [8, 19].

Collecting high-quality, step-by-step vision-language interaction data is expensive and slow; most training corpora are limited to static image-text pairs, producing VLMs that are strong describers but weak actors. Enabling VLMs to acquire practical, transferable skills instead requires methods and data that teach multi-step reasoning and agentic interaction. Multi-turn training in dynamic environments may be the main path to this goal. Simulators offer a practical, scalable alternative, but standard RL algorithms remain fragile: RL4VLM [36] is highly sensitive to a single coefficient balancing "thought" and action loss; LOOP [20] aggregates reward across long trajectories, making credit assignment unreliable in dynamic settings; ArCHer [41] introduces a learned critic, but its

Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), C. Amato, L. Dennis, V. Mascardi, J. Thangarajah (eds.), May 25 – 29, 2026, Paphos, Cyprus. © 2026 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org). https://doi.org/10.65109/

stability depends on dense rewards and large off-policy buffers—both impractical in sparse, long-horizon visual tasks.

Key findings. Experiments across several lightweight simulators—MiniWorld [7], Gym-Cards, ALFWorld [23], and WebShop [33], demonstrate that **transferable agentic skills can be reliably acquired by reinforcement learning (RL) training, provided two conditions are met**: (i) the simulator is inexpensive enough to enable broad exploration, and (ii) the RL algorithm is robust to hyperparameter variation and can be applied out-of-the-box. In this setting, skill transfer is driven by RL training itself; attempts to tune or overfit hyperparameters provide little additional benefit, and, in practice, are a barrier to scaling. Even modest improvements on downstream benchmarks are hard-won, progress on established skill tasks remains nontrivial. Training a VLM in a single, synthetic simulator is sufficient to yield measurable gains on diverse downstream benchmarks. Thus, the main bottleneck is not high-fidelity simulation, but the practicality and hyperparameter robustness of the RL learning rule, a prerequisite for large-scale, reproducible experimentation.

Our approach: Vision-Language Decoupled Actor-Critic (VL-DAC).. To meet that practicality requirement, we propose VL-DAC, an RL objective that cleanly separates the learning signals:

- *Action loss* token-wise Proximal Policy Optimization [22].
- *Value loss* computed once per environment step, with gradients stopped at the VLM backbone.

This token/step split, to our knowledge unused at VLM scale, eliminates RL4VLM's brittle weighting term, avoids LOOP's sequence-level credit-assignment pitfalls, and dispenses with ArCHer's bulky replay buffer and reward requirement. The outcome is a concise, environment-agnostic algorithm that converges faster and ports across simulators with minimal fuss-exactly what is needed to push RL-trained VLMs into new domains at low cost.

Contributions.

- **Vision-Language Decoupled Actor-Critic (VL-DAC)**. We propose an RL objective that pairs token-wise PPO updates with a step-level value head whose gradients are stopped at the VLM backbone; a minimal stabilization kit (KL regularization, value warm-up, and stop-gradient) lets VL-DAC train without the fragile weighting terms or replay buffers required by earlier methods.
- **Skill transfer from synthetic to targeted tasks.** Training a VLM in *one* lightweight simulator at a time already produces notable improvements on targeted skill benchmarks. This highlights the importance of simulator accessibility and algorithmic simplicity for agentic skill acquisition.
- **Skill-transfer analysis.** We present the first systematic evaluation of how simulator-learned capabilities transfer to agentic, spatial, and web-interaction tasks, and we ablate each VL-DAC component to isolate the drivers of stability and generalization.

Overall, our results show that practical and robust RL training, combined with low-cost simulators, unlocks new paths for learning transferable skills in VLMs. VL-DAC provides a step toward environment scaling and broad skill generalization with minimal reliance on hyperparameter tuning or brittle engineering.

2 BACKGROUND

2.1 Vision-Language Agents in Interactive Environments

We model each episode as a finite-horizon Markov Decision Process (MDP) $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma \rangle$, where $\gamma \in [0, 1)$ is the discount factor. Unlike classical RL, the *state* $s_t \in \mathcal{S}$ is a tuple $(\mathbf{x}_t, \mathbf{c}_t)$ consisting of an RGB image (or stack of images) $\mathbf{x}_t \in \mathbb{R}^{H \times W \times 3}$ and an optional text context $\mathbf{c}_t$ (system prompt, dialogue history, etc.).

The *action* $a_t \in \mathcal{A}$ is a sequence of natural-language tokens that fully specifies the next low-level step in the environment (e.g., `"turn_left 15"` or `"click_button id=OK"`).

An agent executes a trajectory $\tau = (s_1, a_1, \ldots, s_T, a_T)$ and seeks to maximize the discounted return

$$J(\theta) = \mathbb{E}_{\tau \sim \pi_\theta} \Big[\sum_{t=1}^{T} \gamma^{t-1} \mathcal{R}(s_t, a_t) \Big],$$

where the *policy* $\pi_\theta(a_t \mid s_t)$ is parameterized by a large vision-language model (VLM) and factorizes auto-regressively,

$$\pi_\theta(a_t \mid s_t) = \prod_{i=1}^{|a_t|} \pi_\theta\big(a_t^{(i)} \mid s_t, a_t^{(<i)}\big).$$

During training, we may additionally learn a state-value function $V_\phi(s_t) = \mathbb{E}_{\tau \sim \pi_\theta}[\sum_{j \geq 0} \gamma^j \mathcal{R}(s_{t+j}, a_{t+j})]$, but the way action and value updates interact differs across methods, as reviewed next. In VL-DAC, we retain this shared backbone but prevent value-head gradients from flowing back, thereby eliminating cross-signal interference.

2.2 Existing RL Algorithms for Multi-Step VLMs & LLMs

Below, we summarize the three baselines that dominate recent work and pinpoint the specific pain points that motivate our *Vision-Language Decoupled Actor-Critic* (VL-DAC) objective introduced in Section 3.

RL4VLM [36]. The policy is decomposed into a "thought" segment (a^{thought}) and an "action" segment (a^{action}). RL4VLM multiplies token-logits of the thought span by $\lambda \in [0, 1]$, effectively rescaling gradient magnitudes:

$$\begin{aligned} &\log \pi_\theta(a_t \mid s_t) = \\ &= \lambda \log \pi_\theta(a_t^{\text{thought}} \mid s_t) + \log \pi_\theta(a_t^{\text{action}} \mid s_t, a_t^{\text{thought}}), \end{aligned} \tag{1}$$

after which, PPO updates are applied at the *step* level. But λ needs to be tuned for each model-environment setup. This makes it hard to scale the method beyond a single environment and limits environment scaling.

LOOP [4]. LOOP employs leave-one-out advantage estimation and trains an LLM in a multi-step scenario using PPO. Because it uses PPO, different policy-update levels (token, step, and trajectory) can be explored; the authors show that the best quality is achieved at the token level. LOO advantage estimation:

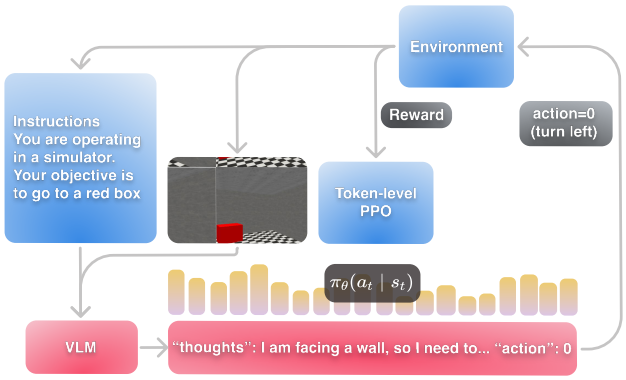

Figure 2: Vision-Language Decoupled Actor-Critic (VL-DAC) pipeline. A vision-language model receives RGB frames and text context, predicts token-wise actions via PPO, and learns a step-level value head whose gradients are stopped at the backbone.

$$A = \frac{K}{K-1}\left(R(s_{0:T}, a_{0:T}) - \frac{1}{K}\sum_{j\neq k} R(s_{0:T}, a_{0:T})\right). \tag{2}$$

The approach sidesteps any need for tuning token mixtures but suffers from extreme credit-assignment noise: a single bad token can wipe out the reward signal for the entire chain, making long-horizon tasks hard to learn.

ArCHer [41]. ArCHer trains a critic with bootstrapped one-step TD [25] at the step level and trains the actor LM from critic feedback. Since the method is primarily designed to be off-policy, it requires a large replay buffer. The method works under *dense* rewards, but two practical issues emerge when we want to train on-policy (e.g., when it is hard to maintain a large buffer) or have sparse rewards (due to the critic design):

- **Replay bottleneck**. Memory demands grow with episode length, which is acute for vision tasks where each step embeds a high-dimensional image, multiple images, or video.
- **Reward sparsity**. When rewards arrive only at episode termination, the critic's bootstrap targets become nearly constant, offering little learning signal.

3 VISION-LANGUAGE DECOUPLED ACTOR-CRITIC (VL-DAC) TRAINING

VL-DAC preserves the intuitive separation between reasoning and behavior (**thought** and **action**) tokens, as in RL4VLM, but crucially removes the fragile balancing coefficient by moving the policy loss to the **token** level while keeping the value loss at the **step** level. Figure 2 illustrates the overall pipeline.

Token-level policy loss. While token-wise RL objectives have appeared in earlier language and vision-language work ([17], [4]), their combination with step-level value prediction has not been systematically studied in large VLMs or for multi-step, skill-centric tasks. Our key innovation is to pair a token-level PPO loss with step-level value targets, resulting in a decoupled objective that provides both learning stability and hyperparameter robustness, a necessity for reliable transfer and scaling. Let $a_t = (a_t^1, \ldots, a_t^{|a_t|})$ be the action tokens at time t. We apply the PPO loss independently to each token:

$$\mathcal{L}_{\text{policy}}^{\text{VL-DAC}}(\theta) = -\mathbb{E}_\tau\left[\frac{1}{|a_t|}\sum_{i=1}^{|a_t|}\min\Big(r_{t,i}A_t, \text{clip}(r_{t,i}, 1-\epsilon, 1+\epsilon)A_t\Big)\right], \tag{3}$$

where $r_{t,i} = \pi_\theta(a_t^i \mid s_t, a_t^{<i}) / \pi_{\theta_{\text{old}}}(a_t^i \mid s_t, a_t^{<i})$ and the advantage A_t is still computed at the *step* level using GAE [21].

Step-level value loss. The value function V_ϕ shares the vision-language backbone with π_θ but has a separate MLP head. We predict $V_\phi(s_t)$ once per step:

$$V_\phi(s_t) = \text{MLP}_\phi\big(\mathcal{F}_{\text{VLM}}(s_t)\big). \tag{4}$$

The value loss is $\mathcal{L}^{\text{Value}}(\phi) = \frac{1}{2}(V_\phi(s_t) - \hat{R}_t)^2$, using step-level advantage estimation via GAE [21].

Stabilization. To stabilize training, we adapt classical RL techniques to the VLM setting: we warm up the value head ϕ for n epochs before updating θ, block gradients from the value head into the backbone (`StopGrad`), and include a per-token forward KL penalty:

$$\mathcal{L}^{\text{KL}}(\theta) = \mathbb{D}_{\text{KL}}\big(\pi_\theta(\cdot \mid s_t) \,\|\, \pi_{\text{old}}(\cdot \mid s_t)\big). \tag{5}$$

Full objective. The total loss is a sum of three terms:

$$\mathcal{L}(\theta, \phi) = \mathcal{L}_{\text{policy}}^{\text{VL-DAC}}(\theta) + \beta\,\mathcal{L}^{\text{KL}}(\theta) + \alpha\,\mathcal{L}^{\text{Value}}(\phi). \tag{6}$$

Empirically, this decoupled approach delivers more stable learning and higher final returns than both RL4VLM and LOOP, and crucially, it enables skill transfer across benchmarks *without fragile hyperparameter tuning* on top of standard PPO parameters. This hyperparameter robustness is vital for scaling RL training to new environments and skills, as tuning-sensitive objectives break down under distribution shift or large-scale experimentation. For details on our prompting setup, see Appendix A.

4 EXPERIMENTS

Our experiments are designed to address four central questions:

- **Q1 Hyperparameter robustness:** Does VL-DAC enable simpler, more robust training than RL4VLM[36] across diverse simulators? We also analyze the contribution of each stabilization technique (KL penalty, value warm-up, stop-gradient), the sensitivity of RL4VLM's λ coefficient (beyond prior work), and performance across different model sizes and architectures.
- **Q2 Credit assignment:** How does VL-DAC compare to LOOP when long-horizon credit assignment is required?
- **Q3 Skill transfer:** Do policies learned in a single, inexpensive simulator transfer agentic skills to downstream benchmarks?

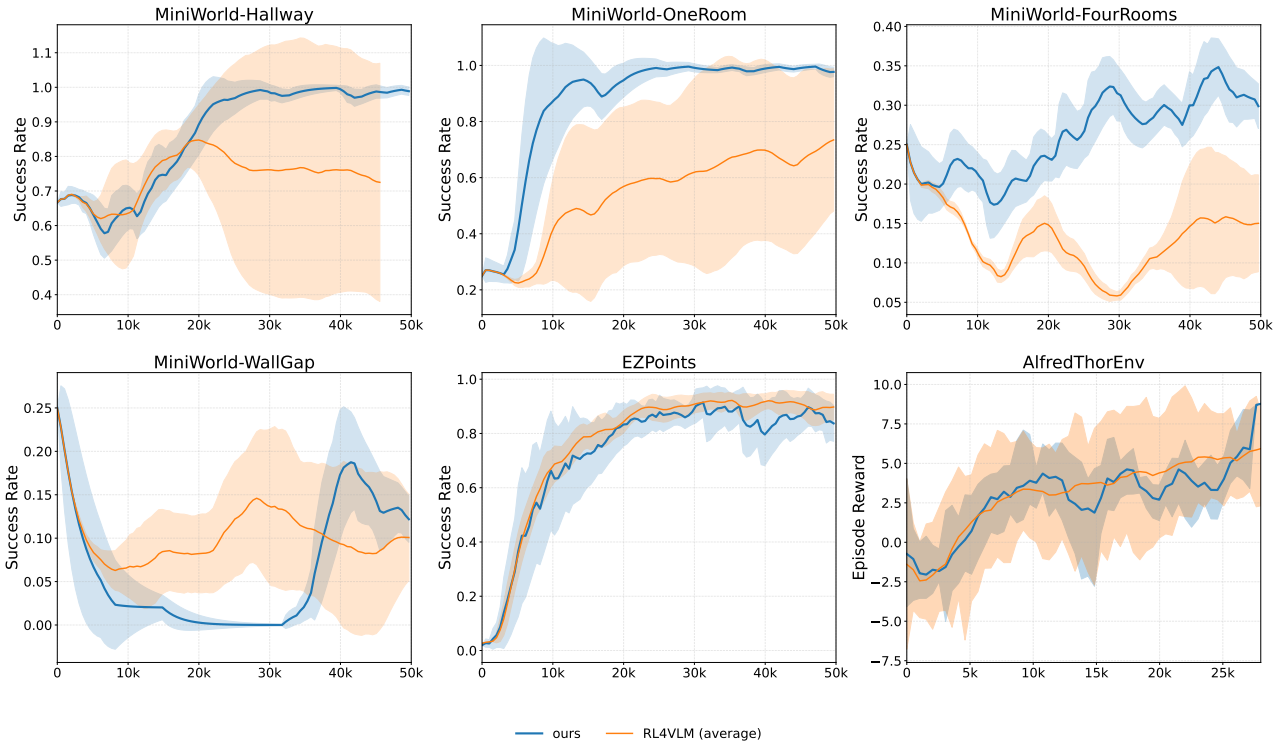

Figure 3: Episode success rates/rewards across environments. Success rates/rewards (%) of our method vs. RL4VLM (averaged over multiple λ values) on six environments: MiniWorld-Hallway, OneRoom, FourRooms (top row), WallGap, EZPoints, ALFWorld (bottom row). While RL4VLM requires tuning λ per environment, our method performs robustly without extra hyperparameter tuning.

Q4 Scalability: Is the method scalable to tasks that require long-term planning and specialized domain understanding (e.g., WebShop), and how does such training impact web-based benchmarks?

We first summarize the experimental setup, then tackle the four questions in turn. We do not include ArCHer in the main-text experiments, since it works poorly under the on-policy scenario (where the training buffer equals the replay buffer) and when rewards are sparse (due to the one-step TD bootstrap). For experiments with ArCHer, see Appendix E.

4.1 Setup

Simulators. We use several lightweight environments: **MiniWorld** (four navigation tasks) for navigation and route-planning, **Gym-Cards/EZPoints** (card-selection logic) as an easy-to-check environment, **ALFWorld** (text-conditioned household tasks) for navigation, spatial reasoning, and agentic capabilities, and **WebShop** (e-commerce browsing) as a domain that requires long-term understanding and web-based planning. All produce RGB frames plus a textual instruction; the agent answers with free-form text that consists of thoughts and action tokens. The total response is parsed into environment actions.

Model and training. We finetune `Qwen2-VL-7B` [29] with LoRA [13] adapters for 25k–50k environment steps, unless otherwise noted. In tables, `base` denotes `Qwen2-VL-7B`. See Appendix B for the hyperparameter grid.

Evaluation metrics. *Simulator success rate* (SR) is the percentage of episodes that reach the goal. *Skill transfer* is assessed using skill-based benchmarks (and their subsets), along with a suite of captioning tasks to check for regressions. For the full evaluation setup, see Appendix C.

Compute budget. Training VL-DAC for 50k environment steps on `Qwen2-VL-7B` takes **20 GPU-hours** on a single NVIDIA H100-80GB. For batch size and other reproducibility parameters see Appendix B.

4.2 Q1. Stability: VL-DAC vs. RL4VLM

Comparison with RL4VLM.. Figure 3 plots SR over 50k steps for *Hallway*, *FourRooms*, *OneRoom*, *WallGap*, *ALFWorld*, and *GymCards*. Curves for RL4VLM are shown as an average of the thought-coefficient λ values recommended by the authors; VL-DAC uses the same optimizer and other hyperparameters, with no extra tuning. VL-DAC reaches high SR in five of six tasks, whereas RL4VLM diverges or plateaus whenever λ is not properly tuned. All RL4VLM experiments here use the same stabilization techniques as VL-DAC. For results without average, performance of top λ per environment and additional details on runs, see Appendix D.

Model	Setup	SR
Qwen2-VL-7B	RL4VLM ($\lambda = 0.35$)	0.98 ± 0.00
Qwen2-VL-7B	RL4VLM ($\lambda = 0.5$)	0.93 ± 0.07
Qwen2-VL-7B	Our	0.98 ± 0.02
Gemma3-4B	RL4VLM ($\lambda = 0.35$)	$0.55 \pm \mathbf{0.38}$
Gemma3-4B	RL4VLM ($\lambda = 0.5$)	$0.82 \pm \mathbf{0.14}$
Gemma3-4B	Our	0.93 ± 0.05

Table 1: RL4VLM vs. VL-DAC. Quality of Qwen2-VL and Gemma over four seeds with varying λ. Qwen2-VL peaks at λ=0.35 in *OneRoom*, while Gemma performs best at λ=0.5. VL-DAC achieves robust, low-variance results across both, even on the more challenging Gemma3-4B. Standard deviations for RL4VLM on Gemma (bold red) highlight severe instability.

Stabilization ablation. Figure 4 shows SR on *OneRoom* when we add KL regularization, value warm-up, and stop-gradient one at a time on top of RL4VLM (λ=0.3, the best setting for *OneRoom* in our experiments). Each component improves convergence speed and reduces variance; all three together boost convergence, and replacing the step-level policy loss with VL-DAC's token-level objective further increases training stability and final quality. The illustrated standard deviation intervals were obtained with four different seeds.

Model and λ comparison. Table 1 summarizes RL4VLM peak SR and standard deviation across λ values and model architectures, alongside VL-DAC's out-of-the-box results. Each entry aggregates four random seeds per setting. For RL4VLM, both the optimal SR and the variability across seeds depend heavily on λ and the model: for example, Gemma3-4B [26] exhibits extremely high standard deviation (**0.38** and **0.14**), regardless of λ, which undermines reproducibility and reliability. Notably, Table 1 also shows that the best-performing λ differs between models, even on the same task—further highlighting that optimal λ values are not stable or transferable across architectures. In contrast, VL-DAC consistently yields strong, low-variance results across models and settings without the need for delicate hyperparameter tuning.

Bottom line. VL-DAC inherits the best of RL4VLM after the stabilization tweaks *and* removes the hyperparameter that still limits RL4VLM in practice due to the need for tuning.

4.3 Q2. Long-horizon credit: VL-DAC vs. LOOP

We compare VL-DAC and LOOP [4] on four MiniWorld environments (Hallway, FourRooms, OneRoom, WallGap) characterized by sparse rewards and long-horizon credit assignment challenges. As shown in Figure 5, LOOP's success rate plateaus within 15-30k steps, whereas VL-DAC continues to improve throughout training. For LOOP we use the authors' recommended configuration. This difference arises from the underlying credit assignment mechanism: LOOP applies the same sequence-level return to every token (algorithm still updates policy on token-level), resulting in high variance and diminishing learning signal as training progresses in long-horizon tasks. In contrast, VL-DAC leverages a step-level critic

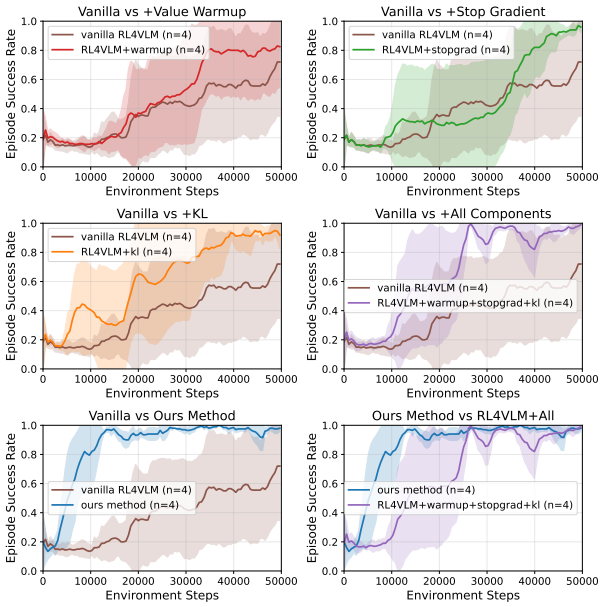

Figure 4: Ablation study of stabilization tricks. Adding KL regularization, value warm-up, and stop-gradient cuts variance sequentially; replacing the step-level policy loss with VL-DAC's token-level objective yields the smooth ascent reported in Figure 3.

	Base	ALFWorld-tuned
$Balrog_{naive}$	$3.21\% \pm 0.75\%$	$\mathbf{4.19\% \pm 0.92\%}$
$Balrog_{CoT}$	$3.94\% \pm 0.98\%$	$\mathbf{6.02\% \pm 1.19\%}$

Table 2: Balrog performance across prompting strategies. RL training (notably VL-DAC) raises scores even with naive prompts, and Chain-of-Thought prompting adds a further >50% boost.

to provide stable, well-shaped advantages at each update, enabling continued learning even when rewards are rare or delayed.

In long-horizon, sparse-reward settings, rewards on sequence-level as in methods like LOOP stall, whereas VL-DAC's decoupled token/step objective continues improving, yielding higher success rate.

4.4 Q3. From MiniWorld/ALFWorld to skill-based benchmark tests

Tables 2 and 4 report downstream skill scores after training in a single simulator at a time.

The *BALROG* benchmark [19] targets long-horizon, agentic skills essential for solving videogame-style tasks, arguably the challenging behaviors to acquire through synthetic training. Table 2 shows that even single-simulator RL training delivers clear improvements on BALROG, with Chain-of-Thought prompting amplifying gains by over 50%. This is particularly remarkable given BALROG's difficulty and the minimal data and compute expended. Other skill

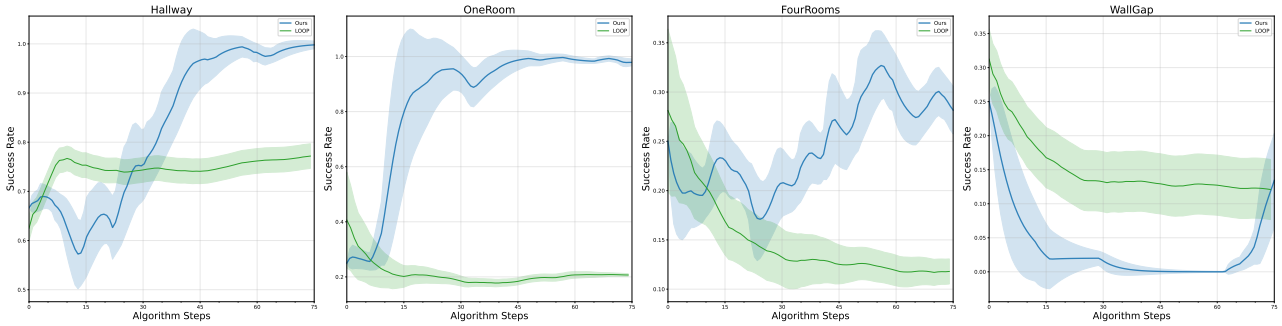

Figure 5: Long-horizon credit assignment: VL-DAC vs. LOOP. On four sparse-reward MiniWorld tasks, LOOP plateaus once early successes exhaust its high-variance sequence-level gradient, whereas VL-DAC continues improving. Token-level updated with step-level advantages and a step-wise critic unlock sustained learning.

benchmarks also show consistent improvement under different training strategies. *VSI*-Bench [32] subsets measure spatial reasoning and planning; *ERQA* [27] assesses spatial understanding; *MuirBench* [28] covers multi-image reasoning; Video-$\text{MME}_{\text{spatial}}$ [11] checks spatial reasoning in video.

- **BALROG:** Training with ALFWorld yields a >50% relative improvement in agentic success, with further gains under Chain-of-Thought prompting. This underscores that multi-step RL environments provide more than surface-level skills,they directly enable the emergence of complex, agentic behavior. Results on BALROG reported over 4 seeds.
- **Skill-specific benchmarks:** ALFWorld training yields a +5% relative gain on the VSI-Bench route-planning task,a type of data that is hard to collect for supervised finetuning, and similar, environment-aligned gains on other subsets. Improvements on ERQA (naive and **CoT** [30]), MuirBench, and Video-$\text{MME}_{\text{spatial}}$ confirm that learned skills generalize across spatial and multimodal reasoning tasks. All skill results are single-seed due to the scale of the datasets.
- **General image and video understanding:** Table 3 shows that general performance is maintained,or even improved, across major vision-language benchmarks ([38], [11], [39], [35], [16], [34], [14], [40], [11], [5], [31]), with largely maintained captioning or perceptual ability after skill-focused RL finetuning.

It is worth noting that prior work with supervised learning often required carefully curated, large-scale datasets to achieve even modest gains on these benchmarks, sometimes with tradeoffs in performance elsewhere. Here, lightweight RL training in one synthetic simulator already delivers measurable, broad-based improvements,most notably on challenging, agentic tasks like BALROG.

4.5 Q4. A different domain: WebShop → VisualWebBench

We next train in **WebShop** for only 2k steps (due to compute constraints). Despite this short training budget, VL-DAC lifts VisualWebBench accuracy by up to +2 percentage points on various subsets compared to the base model, indicating that even limited interaction budgets can yield measurable skill improvements. We also evaluate the impact of models trained on OneRoom on the same benchmark. All results are reported as mean ± standard deviation across 3 seeds. Since some tasks are extraction-based, the mean result across different runs is the same.

Benchmark	Base	ALFWorld	OneRoom	Hallway
GQA	62.02	**62.35**	62.06	62.12
Mirb	**37.38**	36.64	37.25	37.25
$\text{MMBench}_{\text{dev}}$	78.86	78.52	**79.04**	78.52
$\text{MME}_{\text{perception}}$	1681	**1688**	1670	1678
MME-RealWorld	41.81	41.46	41.76	**42.01**
$\text{MMStar}_{\text{avg}}$	56.53	57.03	**57.51**	57.26
$\text{MMT-mi}_{\text{val}}$	59.90	60.40	**60.66**	60.47
MMT_{val}	62.10	62.36	62.65	**62.71**
Video-MME	57.70	**58.11**	57.40	57.70

Table 3: Benchmark gains for Qwen2-VL finetuned on ALFWorld, MiniWorld-Hallway, and MiniWorld-OneRoom. The finetuned model surpasses its instruct baseline in temporal and spatial reasoning, multi-image/video comprehension, and embodied-AI tasks.

5 DISCUSSION

5.1 From a Simple Recipe to a Two-Stage Roadmap

Our results outline a concise two-stage roadmap for converting a vision-language model into a competent interactive agent. **Stage 1** is algorithmic: adopt a token-wise PPO objective with a step-level value head. This decoupling, as realized in VL-DAC, eliminates brittle mixture coefficients, replay buffers, and tuning-sensitive knobs, yielding a *hyperparameter-robust* learner that scales cleanly across model sizes without retuning. **Stage 2** is environmental: expose the agent to lightweight simulators covering diverse action spaces, navigation, manipulation, logic, and browser-based interaction. Stage 1 guarantees a practical, stable RL recipe; Stage 2 provides the behavioral coverage needed for effective skill transfer.

	VSI-Bench$_{\text{route plan}}$	VSI-Bench$_{\text{relative direction}}$	ERQA$_{\text{naive}}$	ERQA$_{\text{CoT}}$	MuirBench	Video-MME$_{\text{spatial}}$
Base	30.93	32.01	38.25	39.00	41.23	64.8
ALFWorld-tuned	**32.47**	31.61	39.00	**39.25**	**42.58**	**66.7**
OneRoom-tuned	31.96	**33.05**	**39.25**	38.50	41.12	**66.7**

Table 4: Skill-specific benchmarks. Models trained in two different environments outperform the base model in their corresponding skill categories.

	web caption	webqa	heading ocr	element ocr	element ground	action prediction	action ground
$base_{naive}$	27.81 ± 0.11	**71.44 ± 0.00**	75.62 ± 1.26	82.36 ± 0.00	**87.49 ± 0.14**	4.98 ± 0.00	83.50 ± 0.00
$base_{cot}$	28.38 ± 0.20	61.11 ± 0.11	74.83 ± 0.00	78.75 ± 0.01	83.29 ± 0.00	6.17 ± 0.21	78.32 ± 0.56
WS_{naive}	**29.31 ± 0.02**	70.32 ± 0.00	**76.34 ± 0.00**	**83.49 ± 0.22**	87.33 ± 0.14	5.34 ± 0.00	82.52 ± 0.00
WS_{cot}	29.04 ± 0.12	62.58 ± 0.05	72.66 ± 0.00	79.95 ± 0.00	84.02 ± 0.00	**6.41 ± 0.00**	78.64 ± 0.00
OR_{naive}	28.19 ± 0.00	70.91 ± 0.00	74.03 ± 0.12	83.31 ± 0.19	86.68 ± 0.00	3.91 ± 0.00	**84.47 ± 0.00**
OR_{cot}	29.21 ± 0.00	59.89 ± 0.00	74.44 ± 0.34	76.20 ± 0.24	83.78 ± 0.00	6.05 ± 0.00	78.64 ± 0.00

Table 5: VisualWebBench breakdown. A short 2k-step WebShop run with VL-DAC improves overall accuracy, with web-captioning and UI-action metrics benefiting most. *WS* refers to WebShop, *OR* to OneRoom.

Importantly, this approach proves especially effective for agentic tasks: achieving measurable improvements on challenging benchmarks such as BALROG highlights the capacity of such approach to increase real decision-making and planning abilities.

5.2 Why Simulator Diversity Matters

Performance gains track the breadth of acquired skills. Training solely on ALFWorld imparts agentic priors that drive a greater than 50% improvement on BALROG, highlighting the challenge and significance of agentic benchmarks. ALFWorld and MiniWorld environments jointly contribute to spatial planning and reasoning, lifting VSI-Bench scores by 5% relative. WebShop, though brief, injects UI-sequencing patterns that yield up to 2% absolute improvement on VisualWebBench. In sum, diverse simulators enable broad, transferable skill acquisition not achievable via single-task or static-image pretraining.

5.3 Limitations and Open Challenges

- **Sparse-reward variance**. Although the critic converges even with terminal rewards, the method still struggles in hard, sparse-reward settings.
- **Beyond screen-based tasks.** All environments studied here involve discrete interface actions on rendered images; continuous-control robotics remains untested.
- **Single-agent assumption**. VL-DAC does not address cooperative or adversarial multi-agent settings where credit must be distributed across agents.
- **Memory and planning**. Current models struggle to process and train in environments that require long-term abstract memory and planning (e.g., MiniWorld-WallGap).
- **Task demands**. Successful training requires models to produce strictly structured, machine-parsable outputs and to maintain coherent chain-of-thought reasoning across steps.

5.4 Future Directions: Scaling the Environment Spectrum

A promising direction is to procedurally generate curricula that expand both task horizons and skill requirements as model capacity increases, similar to MineDojo [10] or Crafter [12] in open-world RL. We envision an open RL4VLM Gym, in which community contributions add *small, inexpensive* environments, fostering a rich spectrum of interactive domains rather than a single monolithic simulation. Such a resource would enable systematic study of *environment-set scaling laws*: for example, how many distinct interaction types are required to achieve an additional *n*% transfer gain?

Another compelling direction is to explore learning across multiple environments simultaneously, allowing the agent to aggregate and transfer knowledge acquired from each domain. This could further enhance generalization and accelerate skill acquisition, compared to sequential or single-environment training.

Algorithmically, VL-DAC could be combined with hierarchical RL, using the step-level value head to supervise subgoal policies while token-wise PPO refines low-level text actions, or paired with memory-augmented transformers to mitigate variance as horizons extend beyond 100 steps.

5.5 Connection to Prior Work

VLM and LLM training in multi-step scenarios. RL4VLM [36], LOOP [4], ArCHer [41], and some other domain-specific methods ([20], [2], [1]) pursue long-horizon training, yet they rely on delicate mixture coefficients, sequence-level gradients with high variance, or replay buffers that collapse under sparse rewards. VL-DAC inherits the stability of PPO-based RLHF while, for the first time, demonstrating *consistent transfer* across agentic, spatial, and web-interaction tasks using the *same* hyperparameters. These findings underscore that a minimal algorithmic tweak, coupled with a

diversified simulator set, is sufficient to unlock practical RL training for VLMs and to endow them with real-world competence.

Benchmarking. Classical perception-centric suites such as MM-Bench, MME, and Video-MME are indispensable for gauging static understanding, but they lack the *agentic* dimension, a capacity to decide and act under long-horizon feedback. Recent game-based evaluations like **BALROG** [19] and **VideoGameBench** [37] close this gap by measuring whether models can plan, execute, and adapt inside fully interactive worlds that resemble classic reinforcement-learning settings. Our study leverages both families: the perception benchmarks verify that VL-DAC training leaves core recognition intact, whereas BALROG [19] exposes the gains in goal-directed control. The contrast underscores a key takeaway: **agentic evaluation is where progress now moves fastest**, and RL with brittle hyperparameters can translate simulator experience into measurable improvements on these harder benchmarks.

Real-task transfer. Generalization from synthetic practice to real-world queries has been actively explored in *single-step* reasoning research ([3], [24]). Our findings extend that evidence to the *multi-step* regime: VL-DAC-trained VLMs master spatial-navigation, manipulation, and web-interaction skills in cheap simulators and then transfer them to BALROG [19], VSI-Bench [32], and VisualWebBench [15] with only modest domain gaps. By showing that interactive rehearsal scales beyond toy boards and text puzzles to full visual control loops, we strengthen the emerging view that *procedural curricula plus lightweight RL* offer a practical path toward robust real-task competence.

6 CONCLUSION

This work demonstrates that reinforcement learning in synthetic, interactive environments is a powerful and scalable approach for advancing vision-language models. By moving from coupled action-critic optimization to a decoupled, two-level objective, and introducing stabilization techniques,we substantially improve the stability and generalization of RL-based training for VLMs. Our method avoids brittle hyperparameter tuning while achieving strong success rates across diverse environments. Crucially, we show that models trained in these synthetic settings generalize effectively to both skill-specific and general-purpose benchmarks,outperforming strong baselines without requiring additional supervision. These results position RL as a viable, data-efficient alternative to traditional supervised fine-tuning, opening new directions for training embodied, multimodal agents capable of reasoning and acting in complex visual domains. Future work will explore scaling to more realistic 3D worlds and incorporating longer-horizon planning into vision-language model training.

REFERENCES

[1] Hao Bai, Yifei Zhou, Mert Cemri, Jiayi Pan, Alane Suhr, Sergey Levine, and Aviral Kumar. 2024. DigiRL: Training In-The-Wild Device-Control Agents with Autonomous Reinforcement Learning. arXiv:2406.11896 [cs.LG] https://arxiv.org/abs/2406.11896

[2] Hao Bai, Yifei Zhou, Li Erran Li, Sergey Levine, and Aviral Kumar. 2025. Digi-Q: Learning Q-Value Functions for Training Device-Control Agents. arXiv:2502.15760 [cs.LG] https://arxiv.org/abs/2502.15760

[3] Jiangjie Chen, Qianyu He, Siyu Yuan, Aili Chen, Zhicheng Cai, Weinan Dai, Hongli Yu, Qiying Yu, Xuefeng Li, Jiaze Chen, Hao Zhou, and Mingxuan Wang. 2025. Enigmata: Scaling Logical Reasoning in Large Language Models with Synthetic Verifiable Puzzles. arXiv:2505.19914 [cs.CL] https://arxiv.org/abs/2505.19914

[4] Kevin Chen, Marco Cusumano-Towner, Brody Huval, Aleksei Petrenko, Jackson Hamburger, Vladlen Koltun, and Philipp Krähenbühl. 2025. Reinforcement Learning for Long-Horizon Interactive LLM Agents. arXiv:2502.01600 [cs.LG] https://arxiv.org/abs/2502.01600

[5] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. 2024. Are We on the Right Way for Evaluating Large Vision-Language Models? arXiv:2403.20330 [cs.CV] https://arxiv.org/abs/2403.20330

[6] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. 2024. Intern VL: Scaling up Vision Foundation Models and Aligning for Generic Visual-Linguistic Tasks. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 24185–24198. https://doi.org/10.1109/cvpr52733.2024.02283

[7] Maxime Chevalier-Boisvert, Bolun Dai, Mark Towers, Rodrigo de Lazcano, Lucas Willems, Salem Lahlou, Suman Pal, Pablo Samuel Castro, and Jordan Terry. 2023. Minigrid & Miniworld: Modular & Customizable Reinforcement Learning Environments for Goal-Oriented Tasks. *CoRR* abs/2306.13831 (2023).

[8] Wei Chow, Jiageng Mao, Boyi Li, Daniel Seita, Vitor Guizilini, and Yue Wang. 2025. PhysBench: Benchmarking and Enhancing Vision-Language Models for Physical World Understanding. arXiv:2501.16411 [cs.CV] https://arxiv.org/abs/2501.16411

[9] DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. 2025. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. arXiv:2501.12948 [cs.CL] https://arxiv.org/abs/2501.12948

[10] Linxi Fan, Guanzhi Wang, Yunfan Jiang, Ajay Mandlekar, Yuncong Yang, Haoyi Zhu, Andrew Tang, De-An Huang, Yuke Zhu, and Anima Anandkumar. 2022. Minedojo: Building open-ended embodied agents with internet-scale knowledge. *Advances in Neural Information Processing Systems* 35 (2022), 18343–18362.

[11] Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, Peixian Chen, Yanwei Li, Shaohui Lin, Sirui Zhao, Ke Li, Tong Xu, Xiawu Zheng, Enhong Chen, Rongrong Ji, and Xing Sun. 2024. Video-MME: The First-Ever Comprehensive Evaluation Benchmark of Multi-modal LLMs in Video Analysis. arXiv:2405.21075 [cs.CV] https://arxiv.org/abs/2405.21075

[12] Danijar Hafner. 2022. Benchmarking the Spectrum of Agent Capabilities. arXiv:2109.06780 [cs.AI] https://arxiv.org/abs/2109.06780

[13] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*. https://openreview.net/forum?id=nZeVKeeFYf9

[14] Drew A Hudson and Christopher D Manning. 2019. Gqa: a new dataset for compositional question answering over real-world images. *arXiv preprint arXiv:1902.09506* 3, 8 (2019), 1.

[15] Junpeng Liu, Yifan Song, Bill Yuchen Lin, Wai Lam, Graham Neubig, Yuanzhi Li, and Xiang Yue. 2024. VisualWebBench: How Far Have Multimodal LLMs Evolved in Web Page Understanding and Grounding? arXiv:2404.05955 [cs.CL] https://arxiv.org/abs/2404.05955

[16] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. 2024. *MMBench: Is Your Multi-modal Model an All-Around Player?* Springer Nature Switzerland, 216–233. https://doi.org/10.1007/978-3-031-72658-3_13

[17] Dipendra Misra, John Langford, and Yoav Artzi. 2017. Mapping Instructions and Visual Observations to Actions with Reinforcement Learning. arXiv:1704.08795 [cs.CL] https://arxiv.org/abs/1704.08795

[18] OpenAI, :, Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, Alex Iftimie, Alex Karpenko, Alex Tachard Passos, Alexander Neitz, Alexander Prokofiev, Alexander Wei, Allison Tam, Ally Bennett, Ananya Kumar, Andre Saraiva, Andrea Vallone, Andrew Duberstein, Andrew Kondrich, Andrey Mishchenko, Andy Applebaum, Angela Jiang, Ashvin Nair, Barret Zoph, Behrooz Ghorbani, Ben Rossen, Benjamin Sokolowsky, Boaz Barak, Bob McGrew, Borys Minaiev, Botao Hao, Bowen Baker, Brandon Houghton, Brandon McKinzie, Brydon Eastman, Camillo Lugaresi, Cary Bassin, Cary Hudson, Chak Ming Li, Charles de Bourcy, Chelsea Voss, Chen Shen, Chong Zhang, Chris Koch, Chris Orsinger, Christopher Hesse, Claudia Fischer, Clive Chan, Dan Roberts, Daniel Kappler, Daniel Levy, Daniel Selsam, David Dohan, David Farhi, David Mely, David Robinson, Dimitris Tsipras, Doug Li, Dragos Oprica, Eben Freeman, Eddie Zhang, Edmund Wong, Elizabeth Proehl, Enoch Cheung, Eric Mitchell, Eric Wallace, Erik Ritter, Evan Mays, Fan Wang, Felipe Petroski Such, Filippo Raso, Florencia Leoni, Foivos Tsimpourlas, Francis Song, Fred von Lohmann, Freddie Sulit, Geoff Salmon, Giambattista Parascandolo, Gildas Chabot, Grace Zhao, Greg Brockman, Guillaume Leclerc, Hadi Salman, Haiming Bao, Hao Sheng, Hart Andrin, Hessam Bagherinezhad, Hongyu Ren, Hunter Lightman, Hyung Won Chung, Ian Kivlichan, Ian O'Connell, Ian Osband, Ignasi Clavera Gilaberte, Ilge Akkaya, Ilya Kostrikov, Ilya Sutskever, Irina Kofman, Jakub Pachocki, James Lennon, Jason Wei, Jean Harb, Jerry Twore, Jiacheng Feng, Jiahui Yu, Jiayi Weng, Jie Tang, Jieqi Yu, Joaquin Quiñonero Candela, Joe Palermo, Joel Parish, Johannes Heidecke, John Hallman, John Rizzo, Jonathan Gordon, Jonathan Uesato, Jonathan Ward, Joost Huizinga, Julie Wang, Kai Chen, Kai Xiao, Karan Singhal, Karina Nguyen, Karl Cobbe, Katy Shi, Kayla Wood, Kendra Rimbach, Keren Gu-Lemberg, Kevin Liu, Kevin Lu, Kevin Stone, Kevin Yu, Lama Ahmad, Lauren Yang, Leo Liu, Leon Maksin, Leyton Ho, Liam Fedus, Lilian Weng, Linden Li, Lindsay McCallum, Lindsey Held, Lorenz Kuhn, Lukas Kondraciuk, Lukasz Kaiser, Luke Metz, Madelaine Boyd, Maja Trebacz, Manas Joglekar, Mark Chen, Marko Tintor, Mason Meyer, Matt Jones, Matt Kaufer, Max Schwarzer, Meghan Shah, Mehmet Yatbaz, Melody Y. Guan, Mengyuan Xu, Mengyuan Yan, Mia Glaese, Mianna Chen, Michael Lampe, Michael Malek, Michele Wang, Michelle Fradin, Mike McClay, Mikhail Pavlov, Miles Wang, Mingxuan Wang, Mira Murati, Mo Bavarian, Mostafa Rohaninejad, Nat McAleese, Neil Chowdhury, Neil Chowdhury, Nick Ryder, Nikolas Tezak, Noam Brown, Ofir Nachum, Oleg Boiko, Oleg Murk, Olivia Watkins, Patrick Chao, Paul Ashbourne, Pavel Izmailov, Peter Zhokhov, Rachel Dias, Rahul Arora, Randall Lin, Rapha Gontijo Lopes, Raz Gaon, Reah Miyara, Reimar Leike, Renny Hwang, Rhythm Garg, Robin Brown, Roshan James, Rui Shu, Ryan Cheu, Ryan Greene, Saachi Jain, Sam Altman, Sam Toizer, Sam Toyer, Samuel Miserendino, Sandhini Agarwal, Santiago Hernandez, Sasha Baker, Scott McKinney, Scottie Yan, Shengjia Zhao, Shengli Hu, Shibani Santurkar, Shraman Ray Chaudhuri, Shuyuan Zhang, Siyuan Fu, Spencer Papay, Steph Lin, Suchir Balaji, Suvansh Sanjeev, Szymon Sidor, Tal Broda, Aidan Clark, Tao Wang, Taylor Gordon, Ted Sanders, Tejal Patwardhan, Thibault Sottiaux, Thomas Degry, Thomas Dimson, Tianhao Zheng, Timur Garipov, Tom Stasi, Trapit Bansal, Trevor Creech, Troy Peterson, Tyna Eloundou, Valerie Qi, Vineet Kosaraju, Vinnie Monaco, Vitchyr Pong, Vlad Fomenko, Weiyi Zheng, Wenda Zhou, Wes McCabe, Wojciech Zaremba, Yann Dubois, Yinghai Lu, Yining Chen, Young Cha, Yu Bai, Yuchen He, Yuchen Zhang, Yunyun Wang, Zheng Shao, and Zhuohan Li. 2024. OpenAI o1 System Card. arXiv:2412.16720 [cs.AI] https://arxiv.org/abs/2412.16720

[19] Davide Paglieri, Bartłomiej Cupiał, Samuel Coward, Ulyana Piterbarg, Maciej Wolczyk, Akbir Khan, Eduardo Pignatelli, Łukasz Kuciński, Lerrel Pinto, Rob Fergus, Jakob Nicolaus Foerster, Jack Parker-Holder, and Tim Rocktäschel. 2024. BALROG: Benchmarking Agentic LLM and VLM Reasoning On Games. arXiv:2411.13543 [cs.AI] https://arxiv.org/abs/2411.13543

[20] Pranav Putta, Edmund Mills, Naman Garg, Sumeet Motwani, Chelsea Finn, Divyansh Garg, and Rafael Rafailov. 2024. Agent Q: Advanced Reasoning and Learning for Autonomous AI Agents. arXiv:2408.07199 [cs.AI] https://arxiv.org/abs/2408.07199

[21] John Schulman, Philipp Moritz, Sergey Levine, Michael Jordan, and Pieter Abbeel. 2018. High-Dimensional Continuous Control Using Generalized Advantage Estimation. arXiv:1506.02438 [cs.LG] https://arxiv.org/abs/1506.02438

[22] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal Policy Optimization Algorithms. arXiv:1707.06347 [cs.LG] https://arxiv.org/abs/1707.06347

[23] Mohit Shridhar, Xingdi Yuan, Marc-Alexandre Côté, Yonatan Bisk, Adam Trischler, and Matthew Hausknecht. 2021. ALFWorld: Aligning Text and Embodied Environments for Interactive Learning. arXiv:2010.03768 [cs.CL] https://arxiv.org/abs/2010.03768

[24] Zafir Stojanovski, Oliver Stanley, Joe Sharratt, Richard Jones, Abdulhakeem Adefioye, Jean Kaddour, and Andreas Köpf. 2025. REASONING GYM: Reasoning Environments for Reinforcement Learning with Verifiable Rewards. arXiv:2505.24760 [cs.LG] https://arxiv.org/abs/2505.24760

[25] Richard S Sutton. 1988. Learning to predict by the methods of temporal differences. *Machine learning* 3, 1 (1988), 9–44.

[26] Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Pier Giuseppe Sessa, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, Amélie Héliou, Andrea Tacchetti, Anna Bulanova, Antonia Paterson, Beth Tsai, Bobak Shahriari, Charline Le Lan, Christopher A. Choquette-Choo, Clément Crepy, Daniel Cer, Daphne Ippolito, David Reid, Elena Buchatskaya, Eric Ni, Eric Noland, Geng Yan, George Tucker, George-Christian Muraru, Grigory Rozhdestvenskiy, Henryk Michalewski, Ian Tenney, Ivan Grishchenko, Jacob Austin, James Keeling, Jane Labanowski, Jean-Baptiste Lespiau, Jeff Stanway, Jenny Brennan, Jeremy Chen, Johan Ferret, Justin Chiu, Justin Mao-Jones, Katherine Lee, Kathy Yu, Katie Millican, Lars Lowe Sjoesund, Lisa Lee, Lucas Dixon, Machel Reid, Maciej Mikuła, Mateo Wirth, Michael Sharman, Nikolai Chinaev, Nithum Thain, Olivier Bachem, Oscar Chang, Oscar Wahltinez, Paige Bailey, Paul Michel, Petko Yotov, Rahma Chaabouni, Ramona Comanescu, Reena Jana, Rohan Anil, Ross McIlroy, Ruibo Liu, Ryan Mullins, Samuel L Smith, Sebastian Borgeaud, Sertan Girgin, Sholto Douglas, Shree Pandya, Siamak Shakeri, Soham De, Ted Klimenko, Tom Hennigan, Vlad Feinberg, Wojciech Stokowiec, Yu hui Chen, Zafarali Ahmed, Zhitao Gong, Tris Warkentin, Ludovic Peran, Minh Giang, Clément Farabet, Oriol Vinyals, Jeff Dean, Koray Kavukcuoglu, Demis Hassabis, Zoubin Ghahramani, Douglas Eck, Joelle Barral, Fernando Pereira, Eli Collins, Armand Joulin, Noah Fiedel, Evan Senter, Alek Andreev, and Kathleen Kenealy. 2024. Gemma: Open Models Based on Gemini Research and Technology. arXiv:2403.08295 [cs.CL] https://arxiv.org/abs/2403.08295

[27] Gemini Robotics Team, Saminda Abeyruwan, Joshua Ainslie, Jean-Baptiste Alayrac, Montserrat Gonzalez Arenas, Travis Armstrong, Ashwin Balakrishna, Robert Baruch, Maria Bauza, Michiel Blokzijl, et al. 2025. Gemini Robotics: Bringing AI into the Physical World. *arXiv preprint arXiv:2503.20020* (2025).

[28] Fei Wang, Xingyu Fu, James Y Huang, Zekun Li, Qin Liu, Xiaogeng Liu, Mingyu Derek Ma, Nan Xu, Wenxuan Zhou, Kai Zhang, et al. 2024. Muirbench: A comprehensive benchmark for robust multi-image understanding. *arXiv preprint arXiv:2406.09411* (2024).

[29] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024. Qwen2-VL: Enhancing Vision-Language Model's Perception of the World at Any Resolution. arXiv:2409.12191 [cs.CV] https://arxiv.org/abs/2409.12191

[30] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems* (New Orleans, LA, USA) *(NIPS '22)*. Curran Associates Inc., Red Hook, NY, USA, Article 1800, 14 pages.

[31] Kai Yan, Zhan Ling, Kang Liu, Yifan Yang, Ting-Han Fan, Lingfeng Shen, Zhengyin Du, and Jiecao Chen. 2025. MIR-Bench: Can Your LLM Recognize Complicated Patterns via Many-Shot In-Context Reasoning? arXiv:2502.09933 [cs.AI] https://arxiv.org/abs/2502.09933

[32] Jihan Yang, Shusheng Yang, Anjali W. Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. 2025. Thinking in Space: How Multimodal Large Language Models See, Remember, and Recall Spaces. arXiv:2412.14171 [cs.CV] https://arxiv.org/abs/2412.14171

[33] Shunyu Yao, Howard Chen, John Yang, and Karthik Narasimhan. 2023. WebShop: Towards Scalable Real-World Web Interaction with Grounded Language Agents. arXiv:2207.01206 [cs.CL] https://arxiv.org/abs/2207.01206

[34] Kaining Ying, Fanqing Meng, Jin Wang, Zhiqian Li, Han Lin, Yue Yang, Hao Zhang, Wenbo Zhang, Yuqi Lin, Shuo Liu, Jiayi Lei, Quanfeng Lu, Runjian Chen, Peng Xu, Renrui Zhang, Haozhe Zhang, Peng Gao, Yali Wang, Yu Qiao, Ping Luo, Kaipeng Zhang, and Wenqi Shao. 2024. MMT-Bench: A Comprehensive Multimodal Benchmark for Evaluating Large Vision-Language Models Towards Multitask AGI. arXiv:2404.16006 [cs.CV] https://arxiv.org/abs/2404.16006

[35] Xiang Yue, Yuansheng Ni, Tianyu Zheng, Kai Zhang, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. 2024. MMMU: A Massive Multi-Discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 9556–9567. https://doi.org/10.1109/cvpr52733.2024.00913

[36] Yuexiang Zhai, Hao Bai, Zipeng Lin, Jiayi Pan, Shengbang Tong, Yifei Zhou, Alane Suhr, Saining Xie, Yann LeCun, Yi Ma, and Sergey Levine. 2024. Fine-Tuning Large Vision-Language Models as Decision-Making Agents via Reinforcement Learning. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*. https://openreview.net/forum?id=nBjmMF2IZU

[37] Alex L. Zhang, Thomas L. Griffiths, Karthik R. Narasimhan, and Ofir Press. 2025. VideoGameBench: Can Vision-Language Models complete popular video games? arXiv:2505.18134 [cs.AI] https://arxiv.org/abs/2505.18134

[38] Kaichen Zhang, Bo Li, Peiyuan Zhang, Fanyi Pu, Joshua Adrian Cahyono, Kairui Hu, Shuai Liu, Yuanhan Zhang, Jingkang Yang, Chunyuan Li, and Ziwei Liu. 2024. LMMs-Eval: Reality Check on the Evaluation of Large Multimodal Models. arXiv:2407.12772 [cs.CL] https://arxiv.org/abs/2407.12772

[39] Yi-Fan Zhang, Huanyu Zhang, Haochen Tian, Chaoyou Fu, Shuangqing Zhang, Junfei Wu, Feng Li, Kun Wang, Qingsong Wen, Zhang Zhang, Liang Wang, Rong Jin, and Tieniu Tan. 2024. MME-RealWorld: Could Your Multimodal LLM Challenge High-Resolution Real-World Scenarios that are Difficult for Humans? arXiv:2408.13257 [cs.CV] https://arxiv.org/abs/2408.13257

[40] Bingchen Zhao, Yongshuo Zong, Letian Zhang, and Timothy Hospedales. 2024. Benchmarking multi-image understanding in vision and language models: Perception, knowledge, reasoning, and multi-hop reasoning. *arXiv preprint arXiv:2406.12742* (2024).

[41] Yifei Zhou and Andrea Zanette. 2024. ArCHer: training language model agents via hierarchical multi-turn RL. In *Proceedings of the 41st International Conference on Machine Learning* (Vienna, Austria) *(ICML'24)*. JMLR.org, Article 2574, 32 pages.