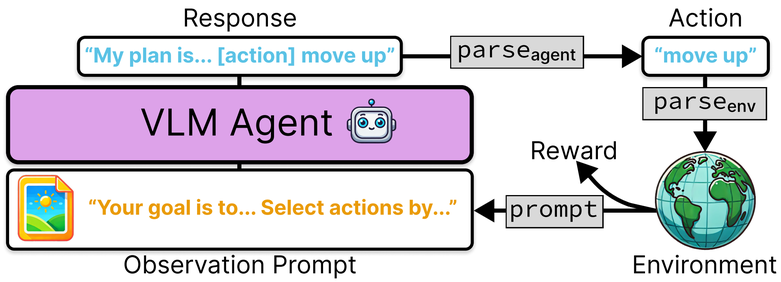
VLM Q-Learning:
Aligning Vision-Language Models for Interactive Decision-Making

Jake Grigsby, Yuke Zhu, Michael Ryoo, Juan Carlos Niebles

VLM Agents

Problem Statement: Open-weight vision-language models lag behind LLMs in agent tasks, struggling with strict output syntax and long-horizon decision-making. Careful prompting and supervised fine-tuning (SFT) improve task alignment, but SFT cannot surpass the quality of our demonstrations, limiting agent performance.

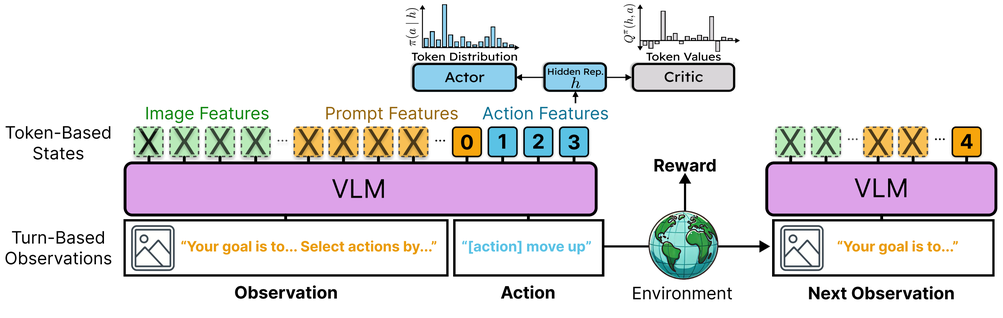
Key Idea: We replace SFT with a simple off-policy RL update: Advantage-Filtered SFT (AFSFT). AFSFT retains the stability of SFT while enabling VLM agents to self-improve from suboptimal datasets by filtering out low-value tokens during training.



Fine-Tuning with Off-Policy Reinforcement Learning

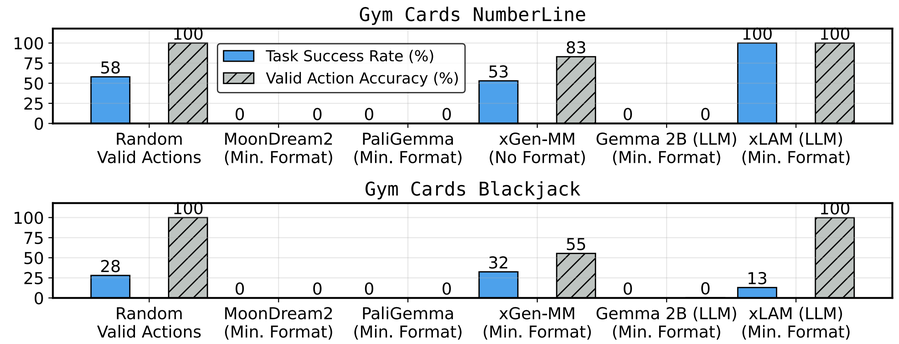
VLM policies map multimodal observations to text actions. We add a critic head to estimate the expected future rewards for each output token. During fine-tuning, we only imitate outputs that we estimate will improve upon our current performance.

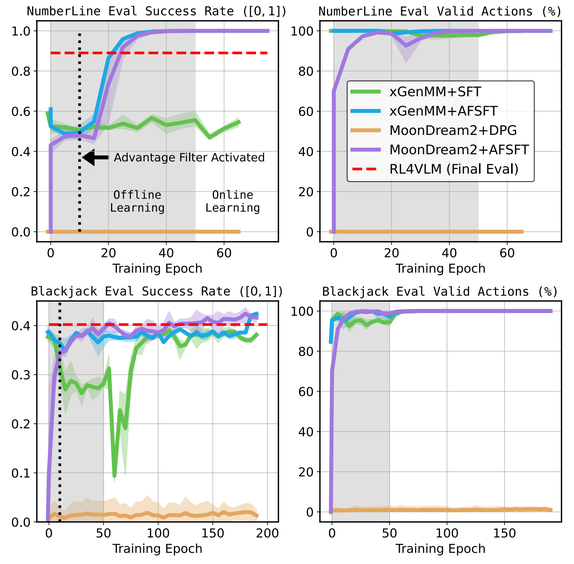
- Temporal difference value learning helps mask suboptimal tokens and learn from inexpensive datasets.
- **Lightweight implementation**: LoRA adapters + a second output head.
- Offline-to-online transition: agent can continue training on its own interactions.
- **Drop-in replacement for SFT:** easily reduces to SFT, and is compatible with any prompting scheme



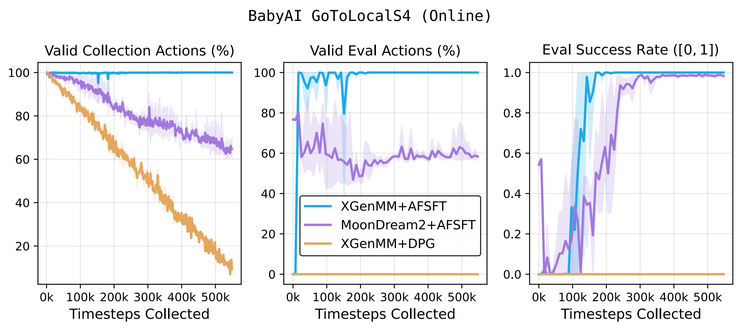
Results

Small base VLMs need task-specific training to produce valid actions. Without RL fine-tuning, prompting can fail to encourage correct syntax in basic tasks →

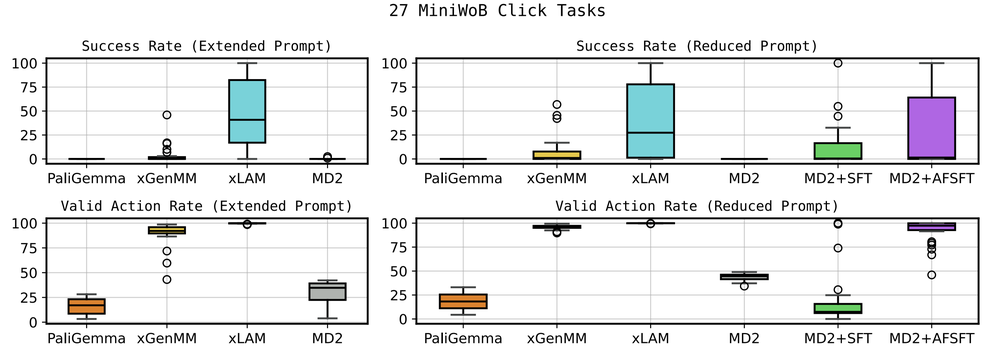




RL fine-tuning aligns syntax and enables self-improvement. Offline AFSFT matches RL4VLM baselines by training on suboptimal (random) actions, while online training in BabyAI shows that agents find valid syntax and successful behavior by training on their own trajectories.



Token-level value learning leverages noisy datasets in open-ended action spaces. In MiniWoB++ click tasks, AFSFT enables VLMs to succeed where SFT fails, improving valid action rates and task success even when training from noisy data collected by prompting base models to interact with a code-like action space.



Scan the QR code for the full paper: