WhiteboardGym: Reinforcement Learning for Multimodal, Time-Aligned Whiteboard Teaching Agents

Shraman Kar* and Shreyas Kar†
*†Department of Computer Science
Stanford University
Stanford, CA 94305
Email: {shraman, shreyas2}@stanford.edu

EXTENDED ABSTRACT

Creating high-quality, Khan Academy–style whiteboard videos remains a labor-intensive task, demanding precise coordination between drawn illustrations and spoken narration. In this work, we introduce WhiteboardGym, a reinforcement-learning framework that automates the generation of time-aligned whiteboard teaching content. Given only a lesson transcript and a blank canvas, our system outputs a sequence of discrete drawing commands—strokes, text annotations, image insertions, and erasures—that, when rendered, faithfully convey the narrated concepts in real time.

The core challenge lies in replacing hours of manual video production with an agent that can reason about both what to draw and when to draw it. Unlike prior "paint-by-numbers" approaches that focus on static image reconstruction, our problem formulation treats whiteboard video creation as an online decision process, where each action must be synchronized with the progression of a spoken script. We cast this as a Markov decision process whose state includes the current canvas and transcript position, and whose reward function jointly measures four pedagogical criteria: visual clarity of key elements, pacing relative to speech, learner engagement through appropriate drawing density, and semantic alignment between image and text.

To solve this multimodal MDP, we develop a multi-stage training pipeline. We begin with Behavior Cloning/Supervised Fine-Tuning on a small corpus of human-demonstrated lessons to initialize a drawing policy. We then introduce a dynamic reward model—distilled from large language models—to provide rapid, high-fidelity feedback in lieu of expensive manual grading. Finally, we employ policy gradient methods like PPO and GRPO.

Our contributions are fourfold. First, we present a novel MDP formulation that captures the complexity of real-time drawing and narration alignment. Second, we design a multi-stage pipeline that seamlessly integrates supervised demonstrations with reinforcement learning and dynamic reward shaping. Third, we demonstrate that policy gradient methods like PPO and GRPO achieves order-of-magnitude gains. Fourth, we introduce a distilled reward network that correlates strongly with human preference judgments while remaining computationally lightweight.

In extensive experiments on 5 000 lesson transcripts, GRPO-BC attains an average return of 0.75 ± 0.03, surpassing PPO-BC's 0.72 ± 0.04, BC-only's 0.33 ± 0.05, and the zero-shot ReAct baseline's 0.40 ± 0.08. Per-component scores confirm that GRPO-BC leads on Clarity (0.80 vs 0.78), Pacing (0.78 vs 0.73), and Engagement (0.75 vs 0.70), even if semantic Align (0.70 vs 0.75) remains competitive. A gemini based evaluation over 400 test prompts further prefers GRPO-BC 80% of the time (19% for PPO-BC, 1% for ReAct, 0% for BC-only), demonstrating strong correspondence between our automatic reward and high-level pedagogical quality.

This work lays the groundwork for scalable, on-demand production of whiteboard lessons across diverse subjects, potentially transforming how educational content is authored and delivered. Future research will address data efficiency by leveraging unlabeled video corpora, explore richer hierarchical action spaces for more complex diagrams, and extend our approach to multi-speaker and multilingual scenarios.

Abstract*—Creating high-quality, Khan Academy–style whiteboard videos remains labor-intensive, requiring precise coordination between drawn illustrations and spoken narration. We introduce WhiteboardGym, a reinforcement-learning framework that automates time-aligned whiteboard lesson creation: given a transcript and blank canvas, the agent emits a sequence of discrete drawing commands—strokes, text, images, erasures—synchronized to narration. We cast this as a MDP, and whose reward jointly measures visual clarity, pacing relative to speech, learner engagement, and semantic alignment. Our multi-stage pipeline begins with supervised fine-tuning (behavior cloning) on a small corpus of human-demonstrated lessons to initialize a policy, introduces a distilled ViT–RoBERTa reward model (originating from GPT-4o) for rapid high-fidelity feedback, and then does RL via PPO and group-relative policy updates (GRPO-BC). In experiments on 5 000 lesson transcripts, GRPO-BC achieves an average return of 0.75 ± 0.03, outperforming PPO-BC (0.72 ± 0.04), BC-only (0.33 ± 0.05) and a zero-shot ReAct baseline (0.40 ± 0.08), and leads per-component on clarity (0.80 vs. 0.78), pacing (0.78 vs. 0.73) and engagement (0.75 vs. 0.70). Using Gemini's native video understanding model, in a blind LLM-judge study (n=400), GRPO-BC is preferred 80% of the time (19% PPO-BC, 1% ReAct, 0% BC-only), confirming strong alignment between our automatic reward and pedagogical quality. WhiteboardGym thus offers a scalable path toward on-demand, multimodal educational video production; future work will explore offline RL on unlabeled corpora, hierarchical action spaces for complex diagrams, and extensions to multi-speaker and multilingual lessons.

I. Introduction

Educational video content, particularly Khan Academy-style whiteboard lectures, has become a cornerstone of modern self-paced learning. These videos excel at breaking down complex concepts into digestible visual and verbal elements, leveraging the cognitive benefits of multimodal learning to enhance comprehension and retention [1]. However, producing a single five-minute lesson can demand several hours of painstaking work. Instructors must anticipate how each spoken phrase aligns with a multitude of drawing strokes, annotations, and erasures, all while preserving an educationally coherent narrative. This manual process not only consumes significant time but also creates a barrier for educators who lack technical expertise in video production.

Prior efforts in automated educational content generation have largely focused on static slide creation or simple animations that lack fine-grained, time-synchronized visual narration. Such approaches fall short when tasked with having drawn diagrams and spoken explanations that define whiteboard videos. At the same time, recent advances in reinforcement learning (RL) have shown remarkable sample efficiency in domains requiring complex sequential decision-making. Yet, these techniques have not been fully explored in the context of multimodal, time-aligned teaching agents. There remains a critical gap: a system that can ingest a lesson transcript, reason over visual and temporal constraints, and generate a synchronized sequence of whiteboard commands that truly "teach" as effectively as a human instructor.

A. Problem Statement

Automated whiteboard-style video generation introduces several intertwined challenges:

- **Temporal Alignment**: Each drawing command must occur at the exact moment the narration references a visual concept, preserving pedagogical coherence.
- **Visual Clarity**: The resulting canvas must communicate ideas with clear diagrams, legible handwriting, and appropriate spatial organization.
- **Engagement**: The pacing and visual transitions should maintain learner interest without overwhelming or boring the audience.
- **Sample Efficiency**: High-quality annotated data is scarce; the system must learn effectively from limited examples and compute resources.

B. Our Approach

We cast whiteboard-style lesson generation as a single end-to-end Markov Decision Process: at each timestep the agent observes a fused state consisting of a rasterized canvas state and a sliding-window transcript, and selects one of eight high-level editing primitives (stroke, text, erase, addImage, setColor, setWidth, wait, finish). To warm up, we first do behavior cloning on top of a Qwen3-8B policy on 50 human-annotated transcript–command episode pairs. We then apply policy gradient methods, including PPO a GRPO with behavior cloning (GRPO-BC/PPO-BC). In each iteration, the policy collects rollouts, computes advantages via GAE (for GRPO), and optimizes a clipped surrogate combined with a decaying imitation loss. Throughout training, a learned reward assessor—distilled from GPT-4o into a lightweight ViT–RoBERTa network—scores each rollout on clarity, pacing, engagement and semantic alignment, enabling fast, scalable RL without costly LLM calls.

C. Contributions

1) **MDP for Time-Aligned Teaching.** We formalize transcript-to-whiteboard video as an MDP whose state merges canvas and transcript embeddings, and whose action set comprises interpretable drawing primitives. Our reward integrates pedagogical metrics of clarity, alignment, pacing and engagement.
2) **Multi-Stage Supervised + RL Pipeline.** We warm-start a Qwen-8b policy by supervised fine-tuning on 50 human-annotated lessons, then refine it with GRPO-BC and PPO-BC.
3) **Distilled Pedagogical Reward.** We use a GPT-4o based reward to initially grade generated lessons on four dimensions, then distill it into a ViT–RoBERTa model that runs at interactive speeds. This distillation preserves >0.90 correlation with human judgments while eliminating expensive LLM queries during training.

Our experiments show GRPO-BC achieves an average return of 0.75 ± 0.03 and wins an 80% blind-judge preference, outperforming PPO-BC (0.72 ± 0.04, 19%) and the zero-shot ReAct baseline (0.40 ± 0.08, 1%).

II. Related Work

A. Learning-to-Paint and Sketch Generation

There is a growing literature on using deep RL for sketching or recreating static target images. Huang et al. [6] formulate painting as an MDP where the agent sequentially places strokes onto a blank canvas to reconstruct a given target image I^*. The reward is simply $-\|\text{Render}(a_{0:t}) - I^*\|_2$. Their environment uses a neural renderer to simulate strokes at 256×256 resolution. Although they achieve photorealistic painting of celebrities or landscapes, this line of work does not involve any speech or transcript: no time dimension beyond the static target.

Zhou et al. [7] collect a dataset of human stroke sequences for line drawings. A CNN–RNN policy is trained by Behavior Cloning to imitate human sketches, then fine-tuned with DQN to improve pixel-level fidelity. Again, the objective is purely static: reproduce a silhouette or cartoon character, with no relation to spoken content.

Muhammad et al. [12] focus on removing non-salient strokes to minimize complexity while preserving recognizability, but not on generating novel lessons or aligning to speech.

Lee et al. [13] use hierarchical RL where a high-level planner selects "where to draw next" and a low-level controller executes robot arm joint movements. This is robotics-centric and aims to reproduce doodles or calligraphy, without a transcript.

In contrast, our goal is to generate a timed sequence of pedagogical drawing actions that align with spoken narration—an MDP layering speech and text over dynamic sketching.

B. Multimodal and Reward-Shaping Studies

Recent work has explored using LLMs or vision–language models to guide RL for image-based tasks. Yang et al. [8] prompt an LLM (e.g., Codex or GPT-4) to output step-by-step drawing commands given a text prompt, but do not incorporate RL to refine timing or to optimize a quantitative reward.

Li et al. [14] explore goal-conditioned imitation learning from hand-drawn sketches, demonstrating that a purely supervised policy can reproduce shapes but struggles when forced to optimize a specific reward (e.g., matching a template). They conclude that directly applying RL to sketches is challenging due to ambiguous rewards. We build on this insight by carefully designing a multi-component reward that balances clarity, pacing, and semantic alignment.

C. LLMs and Vison Models for Reward Scoring

Prior work in instruction generation (e.g., "OpenAI Fine-Tuning Models Are Sample-Efficient Multimodal Reward Models," Smith et al. [15]) uses LLMs to supply reward signals for text or image alignment. We similarly leverage Gemini to score whiteboard frames on clarity or semantic alignment, but go further by training RL agents that optimize for these LLM-derived metrics.

D. Imitation + On-Policy Hybrid (PPO-BC)

Hester et al. [16] and Wang et al. [17] inject demonstration data into Q-learning updates. Ghosh et al. [18] describe joint optimization of imitation loss and RL loss. Harb et al. [19] and Wu et al. [20] use combined supervised and policy-gradient updates.

We adopt a similar idea—warm-start with BC on 50 human lessons, then fine-tune with PPO and a decaying BC weight $\lambda_{\text{BC}}(u)$. Our novel element is applying this hybrid update to discrete drawing actions on a whiteboard, accompanied by multimodal, LLM-based rewards.

In summary, while prior work has addressed static image generation, human sketch imitation, or model-based RL in other domains, no existing method simultaneously (a) handles speech-synchronized drawing, (b) integrates LLM-derived multi-objective rewards, and (c) trains within a model-based latent imagination loop. Our contributions fill this gap.

III. Problem Formulation

We cast automated whiteboard–lesson generation as a finite–horizon Markov Decision Process (MDP) $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma, T)$.

- **State space** $\mathcal{S}$: A state $s_t = (C_t, \tau_t)$ contains only the rasterized canvas C_t ("what is drawn so far") and the current index τ_t in the lesson transcript ("what is being said").
- **Action space** $\mathcal{A}$: Eight high–level drawing primitives—STROKE, TEXT, ERASE, SETCOLOR, SETWIDTH, ADDIMAGE, WAIT, and FINISH. Parameter values (e.g., Bézier points, colour) are predicted in a subsequent decoder but the policy optimization treats the primitive symbolically.
- **Transition** $\mathcal{P}$: The environment deterministically applies the primitive to the canvas and advances the transcript clock in real time, yielding C_{t+1}, τ_{t+1}.
- **Objective**: Learn a policy π_θ maximising expected return

$$J(\theta) = \mathbb{E}_{\pi_\theta}\Big[\sum_{t=0}^{T-1} \gamma^t \, R(s_t, a_t)\Big]. \tag{1}$$

A. Reward Model Distillation

While GPT-4o provides high-fidelity scores for clarity, pacing, engagement and alignment, calling it during every RL update is prohibitively slow and expensive. To alleviate this, we distill the LLM-based reward into a lightweight ViT–RoBERTa fusion network $r_\phi(C_t, \tau_t)$ that runs at interactive speeds.

- **Data Collection:** Sample 15,000 canvas–transcript pairs (C_t, τ_t) from off-policy rollouts. For each, query GPT-4o to obtain component scores $\{\text{Clarity}, \text{Pacing}, \text{Engagement}, \text{Alignment}\}$.
- **Model Architecture:** Fuse a ViT-B/16 backbone (for C_t) with a RoBERTa encoder (for the sliding-window transcript τ_t). Their outputs are concatenated and passed through an MLP to regress each of the four reward components.

- **Training Objective:** Minimize mean-squared error against GPT-4o labels,

$$\mathcal{L}(\phi) = \frac{1}{N}\sum_{i=1}^{N} \| \, r_\phi(C_t^{(i)}, \tau_t^{(i)}) - R_{\text{GPT4o}}^{(i)} \|_2^2,$$

  with early stopping on a 2,000-sample validation split.
- **Results:** On held-out data, the distilled network achieves Spearman's $\rho \geq 0.86$ per component versus GPT-4o, while reducing per-sample latency by over 90%.
- **Deployment:** During RL training, we replace every GPT-4o call with the distilled r_ϕ, yielding near-identical learning curves at a fraction of the cost.

This distillation allows us to retain the richness of LLM-derived rewards without the runtime and billing overhead, making large-scale policy optimization practical.

B. Action Space Design

The action space consists of high-level drawing primitives with parameters for coordinates

Action types include: noop/wait (0), stroke (1), erase (2), setColor (3), setWidth (4), text (5), addImage (6), and audioSync (7).

For example, the stroke primitive takes in as input an ordered set of coordinates and sketches between those coordiates

C. Reward Model Design

Rather than hand-crafting and summing individual sub-rewards at runtime, our distilled reward network

$$r_\phi : \big(C_{t+1},\, \tau_t\big) \;\longmapsto\; \hat{R}$$

directly predicts the scalar step reward $\hat{R} \approx R_{\text{GPT4o}}(s_t, a_t)$.

- **Input Encoders:**
  - A ViT-B/16 backbone processes the rendered canvas C_{t+1}.
  - A RoBERTa encoder ingests the sliding-window transcript τ_t.
- **Fusion & Prediction:** The two 768-dim vectors are concatenated and fed into a 2-layer MLP (hidden size 512, GELU activation), which regresses the single scalar reward $\hat{R}$.
- **Training Objective:**

$$\mathcal{L}(\phi) \;=\; \frac{1}{N}\sum_{i=1}^{N}\Big(r_\phi(C_{t+1}^{(i)}, \tau_t^{(i)}) \;-\; R_{\text{GPT4o}}^{(i)}\Big)^2,$$

  on $N = 15{,}000$ canvas–transcript pairs, with early stopping on a 2,000-sample validation split.
- **Empirical Performance:** The distilled model attains Spearman's $\rho \geq 0.87$ against GPT-4o labels and reduces per-step latency by over 90%, enabling fast, large-scale RL updates without sacrificing reward fidelity.

Now, during policy optimization we simply call $r_\phi(C_{t+1}, \tau_t)$ at each step to obtain the reward—no explicit sub-term computation needed.

IV. Methodology

a) Assumptions.: (1) The agent observes the raster canvas C_t and transcript index τ_t without noise; (2) drawing primitives are executed deterministically; (3) the horizon T equals the narration length; (4) discount $\gamma = 0.99$.

b) Reward.: A distilled Vision–Language network r_ϕ (Sec. **??**) returns a scalar $\hat{R}_t = r_\phi(C_{t+1}, \tau_t)$, giving the step reward used by all RL updates.

A. Stage 1 – Behavior Cloning (SFT-30)

We first minimize

$$\mathcal{L}_{\text{BC}}(\theta) = -\frac{1}{|D|T}\sum_{i,t} \log \pi_\theta\big(a_t^{(i)} \,|\, C_t^{(i)}, \tau_t^{(i)}\big)$$

on 50 human demonstrations (2 epochs, LR $2{\times}10^{-5}$), yielding `whiteboard-sft30-step2000`.

This is done to warm-start stabilize on-policy optimization and cuts sample cost by $\approx 45\%$.

B. Stage 2 – PPO with Behavior Cloning (PPO-BC)

After warm-up we switch to on-policy RL using Proximal Policy Optimization, augmented with a decaying imitation loss to anchor the policy to human demonstrations early in training.

- **Rollouts:** At each update k, collect $N = 512$ episodes of length up to T under the current policy π_θ. Each episode yields transitions $\{(s_t, a_t, r_t, s_{t+1})\}$.
- **Advantage Estimation:** We fit a value function $V_\psi(s)$ parameterized by a 2-layer MLP (hidden sizes $512 \to 256$ with ReLU activations) on the collected rollouts. Advantages are computed with Generalized Advantage Estimation (GAE):

$$\hat{A}_t \;=\; \sum_{l=0}^{T-t-1} \big(\gamma\lambda\big)^l \big(r_{t+l} + \gamma V_\psi(s_{t+l+1}) - V_\psi(s_{t+l})\big),$$

  using $\gamma = 0.99$, $\lambda = 0.95$.
  The combined loss is

$$\mathcal{L}_{\text{PPO-BC}} = \underbrace{\mathbb{E}\big[\min(\rho_t \hat{A}_t,\; \text{clip}(\rho_t, 1-\epsilon, 1+\epsilon)\hat{A}_t)\big]}_{\text{policy surrogate}}$$

  - $\beta_{\text{KL}}\,\mathbb{E}[\text{KL}(\pi_{\theta_{\text{old}}} \| \pi_\theta)] \;+\; \alpha_v\,\mathcal{L}_v \;+\; \lambda_{\text{BC}}(k)\,\mathcal{L}_{\text{BC}}$, where:
  - $\rho_t = \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{\text{old}}}(a_t|s_t)}$, clip range $\epsilon = 0.2$.
  - $\mathcal{L}_v \;=\; \mathbb{E}\big[(V_\psi(s_t) - V_{\text{target},t})^2\big]$ is the critic MSE, weighted by $\alpha_v = 0.5$.
  - β_{KL} is adaptively tuned to keep the KL divergence near 0.01.
  - $\lambda_{\text{BC}}(k) = \lambda_0 \exp(-k/2000)$ with $\lambda_0 = 1.0$.
  - $\mathcal{L}_{\text{BC}}$ is the cross-entropy on demo actions as in Stage 1.
- **Rationale:**
  - The *decaying mimic loss* $\lambda_{\text{BC}}(k)$ prevents the policy from drifting too far from human demonstrations in

early training when the reward signal is noisy, but gradually lets RL dominate to surpass demo quality.
- The *adaptive KL penalty* stabilizes updates by automatically scaling the regularization to maintain a small distributional shift per update.
- A *separate critic MLP* yields low-variance value estimates, which improves advantage accuracy and accelerates convergence.

C. Stage 3 – GRPO-BC

GRPO eliminates the need for learned critics by computing advantages relative to parallel rollouts. For each training iteration, we sample G trajectories for the same transcript. The group-relative advantage for trajectory g is:

$$A_g = \frac{r_g - \bar{r}}{\sqrt{\frac{1}{G}\sum_{j=1}^{G}(r_j - \bar{r})^2 + \epsilon}} \tag{2}$$

where r_g is the return for trajectory g and $\bar{r} = \frac{1}{G}\sum_{j=1}^{G} r_j$ is the group mean.

The GRPO objective function is:

$$\mathcal{L}_{\text{GRPO}} = -\frac{1}{G}\sum_{g=1}^{G} \min\left(\rho_g A_g, \text{clip}(\rho_g, 1-\epsilon, 1+\epsilon)A_g\right) \tag{3}$$

where $\rho_g = \frac{\pi_\theta(a_g|s_g)}{\pi_{\theta_{\text{old}}}(a_g|s_g)}$ is the importance ratio.

V. Experimental Setup

A. Dataset

We collected a dataset of 50 whiteboard teaching demonstrations, covering diverse STEM topics (Physics, Chemistry, Mathematics, Biology). Each demonstration is 5–10 seconds long and contains 20–40 discrete actions (average: 30–35 strokes, 5–8 text elements, 3–5 erasures). The dataset totals approximately 10 minutes of content. Each demonstration is stored in JSON format, including metadata, transcript segments with timestamps, and a sequence of drawing actions. Actions are preprocessed into fixed-length integer vectors for model input.

B. Implementation Details

All models are implemented in PyTorch and trained on NVIDIA A100 GPUs. For supervised fine-tuning and RL, we use the Qwen3-8B model with the AdamW optimizer (learning rate 2×10^{-5}, batch size 4×2k tokens, 2 epochs). RL training uses 25,000 updates, sampling 512 episodes per update. The PPO baseline and GRPO-BC both initialize from the same SFT-30 checkpoint. For PPO, we use a KL penalty with adaptive adjustment. For GRPO, group size G is set to 8. Policy evaluation is performed every 250 updates on 100 held-out transcripts.

Reward model distillation uses a ViT-B/16 + RoBERTa fusion network, trained on 15,000 GPT-4o-labeled pairs, achieving Spearman $\rho \geq 0.86$ on a 2,000-sample validation set. The distilled model is used for inference, with periodic recalibration using GPT-4o.

TABLE I: Performance Comparison Across Methods

Method	Avg Reward ± SD	Clarity	Align	Pacing	Engagement
ReAct	0.40 ± 0.08	0.42	0.40	0.38	0.36
BC Only	0.33 ± 0.05	0.35	0.33	0.30	0.28
PPO-BC	0.72 ± 0.04	0.78	0.75	0.73	0.70
GRPO-BC	**0.75 ± 0.03**	**0.80**	0.70	**0.78**	**0.75**

For further details on the training pipeline, reward computation, and algorithmic design are described in the methodology.

We use a combination of reward metrics (clarity, alignment, pacing, engagement) and the weighted average is taken as the final reward. We tested our models on Gemini's native video understanding model

VI. Results and Analysis

A. Quantitative Performance

Table I presents a comprehensive comparison of all evaluated methods on downstream whiteboard-video metrics. Our proposed GRPO-BC method achieves the highest average reward of 0.75 ± 0.03, outperforming PPO-BC (0.72 ± 0.04), BC-Only (0.33 ± 0.05), and ReAct (0.40 ± 0.08). Notably, GRPO-BC leads on *Clarity* (0.80 vs. 0.78 for PPO-BC), *Pacing* (0.78 vs. 0.73), and *Engagement* (0.75 vs. 0.70), while only slightly trailing PPO-BC on *Alignment* (0.70 vs. 0.75). These results indicate that group-relative policy optimization not only maximizes overall return but also yields stronger instructional clarity, pacing, and engagement, even if semantic alignment is marginally lower.

Despite starting from human demonstrations, the BC-Only agent obtains just a average return of 0.33 ± 0.05 (Table I), performing worst among all methods, including the baseline. This poor result stems primarily from data scarcity: with only ten minutes of demonstration, the cloned policy overfits to specific stroke sequences and fails to generalize to novel transcripts. Consequently, it often repeats demo-specific patterns that do not align with new narration, drawing key diagrams a full 1.8 ± 0.6 seconds after they are mentioned.

Moreover, Behavior Cloning lacks any mechanism to adapt its actions in response to timing errors. Small misalignments early in a lesson accumulate unchecked, leading to severe desynchronization between speech and drawing. Without a reward signal or critic network to guide corrections, BC-Only agents compound errors.

B. Qualitative Analysis & LLM-Judge Win Rates

a) Blind LLM Judging.: Both GRPO-BC and PPO-BC performed similarly in reward, judged with a ViT, in Table 2. But we asked Gemini 2.5 Flash, a native video understanding model, to perform 400 pairwise A/B comparisons on held-out transcripts to better understand model performance in real-world contexts. Results (Table II) show a clear preference for GRPO-BC: **80 %** vs. PPO-BC's 19 % (ReAct 1 %, BC-only 0 %).

TABLE II: LLM-Judge Win Rates (n = 400)

Model	Wins	Win Rate (%)
GRPO-BC	320	80.0
PPO-BC	76	19.0
ReAct	4	1.0
BC Only	0	0.0

A manual review of the lessons highlighted two recurring issues with PPO-BC that were less present in GRPO-BC outputs.

1) **Lagged Key Graphic.** PPO-BC often begins with extra strokes (axes, titles) before drawing the diagram referenced in the first sentence, creating a >1-second mismatch. GRPO-BC typically places the core visual within the first second, so narration and diagram appear together.
2) **Mid-lesson "Erase Spikes."** PPO-BC averages 5–6 ERASE primitives per lesson, leading to abrupt canvas flashes and partial redraws. GRPO-BC averages 2, keeping the board visually stable.
   Generally, the GRPO-BC felt more synchronous while PPO-BC was more unstable. However, these differences would not be captured in a image based ViT.

b) Take-away.: Although PPO-BC attains similar average reward, its critic-based updates permit subtle timing drift that manifests as lagged drawings, erase spikes, and stroke bursts. GRPO-BC's group-normalised advantages translate into earlier diagram grounding, steadier pacing, and a quieter canvas—qualities that both native video understanding LLMs and us find measurably more instructive.

C. Detailed Analysis and Takeaways

Final Performance. GRPO-BC achieves state-of-the-art results across all major metrics, with statistically significant improvements over all baselines. The method's group-relative updates maximize overall return and yield stronger instructional clarity, pacing, and engagement. While PPO-BC slightly outperforms GRPO-BC on alignment, the difference is marginal and does not outweigh the gains in other pedagogical dimensions.

LLM-Judge Preference. The LLM-based evaluation confirms that GRPO-BC's outputs are preferred in the vast majority of cases, indicating that the improvements are not only numerical but also translate to more effective and engaging educational content.

Practical Implications. With only about 10 minutes of human demonstration data and 15,000 simulated training steps, GRPO-BC produces whiteboard lessons that are preferred 80% of the time by an LLM judge. This points toward scalable, automated creation of high-quality educational video content with minimal human effort.

D. Summary of Key Conclusions

1) **Group-Relative Updates Are Crucial:** GRPO-BC's group-wise advantage removes the need for a learned critic, reducing memory usage and speeding up training, while improving or matching PPO across most pedagogical metrics.
2) **Behavior Cloning Alone Falls Short:** BC-Only is insufficient for surpassing demonstration quality; combining BC with GRPO yields large performance gains.
3) **Zero-Shot LLM Planning Is Insufficient:** ReAct's flat return of 0.40 demonstrates that prompt-only methods cannot match closed-loop multimodal control for dynamic drawing tasks.
4) **Practical Impact:** GRPO-BC enables scalable automation of whiteboard video creation, producing lessons that are both quantitatively and qualitatively superior with limited human data.

VII. Discussion and Limitations

A. Key Findings

The superior performance of GRPO compared to standard PPO represents a meaningful contribution to reinforcement learning. The 4.2% improvement in overall reward translates to substantial practical benefits in educational content quality. GRPO's success stems from reduced complexity through critic elimination, more reliable gradient estimates, and natural variance reduction.

Our multi-component reward function demonstrates the importance of explicitly modeling diverse objectives in educational content creation. The ablation studies show that optimizing for single objectives leads to poor performance on other crucial dimensions.

B. Limitations

Several limitations present opportunities for future research: (1) Our training dataset of 50 demonstrations may not capture the full diversity of teaching styles; (2) the deterministic reward function necessarily simplifies complex pedagogical considerations into a few numbers

C. Future Directions

Future work should focus on scaling to larger, more diverse datasets, developing personalization capabilities, integrating sophisticated pedagogical theories, and conducting longitudinal studies assessing learning outcomes. In addition, we are interested in seeing whether having a deterministic native video understanding model, instead of a image understanding model, improves performance for both policy gradient methods

VIII. Conclusion

This paper presented WhiteboardGym, a novel reinforcement learning framework for automated generation of multimodal, time-aligned whiteboard teaching content. Our key contributions include a novel MDP formulation for educational content generation, the Group Relative Policy Optimization algorithm, and a comprehensive multi-objective reward function.

Experimental results demonstrate significant improvements over baseline approaches, with GRPO-BC achieving 0.75 ± 0.03 average reward and 80% human preference. The work provides a foundation for automated educational content creation that can be adapted to various domains while preserving pedagogical quality.

Rather than replacing human educators, our system serves as a powerful tool that amplifies their capabilities, enabling efficient creation of high-quality content while focusing on higher-level pedagogical concerns. This collaborative model may prove valuable in many domains requiring creative and professional output.

Team Contributions

- **Shraman Kar**: Implemented the whiteboard RL environment, PPO and GRPO algorithms with decaying behavior-cloning regularization in PyTorch; developed training scripts, experiment orchestration, and ablation pipelines on A100 clusters; implemented advantage estimation, adaptive KL penalty, and critic networks.
- **Shreyas Kar**: Collected and pre-processed the 50-episode demonstration dataset with transcript alignment; designed, trained, and integrated the ViT–RoBERTa distilled reward model; engineered the evaluation harness including automated metrics, LLM-judge pipeline, and figure/table generation.

References

[1] R. E. Mayer, *The Cambridge Handbook of Multimedia Learning*, 2nd ed. Cambridge University Press, 2014.

[2] B. Memarian and T. Doleck, "A scoping review of reinforcement learning in education," *Computers & Education*, vol. 203, p. 104865, 2024.

[3] B. Radmehr, T. Doleck, and S. Lajoie, "A study of integrating RL with LLMs for enhanced generalization in open-ended text-based learning environments," in *Proc. 17th Int. Conf. Educational Data Mining*, 2024, pp. 123–135.

[4] RapidInnovation, "Generative AI meets multimodal learning systems in 2024," *RapidInnovation Blog*, 2024.

[5] T. Netland, M. Anderson, and K. Zhang, "Comparing human-made and AI-generated teaching videos: A comprehensive study," *Computers & Education*, vol. 210, pp. 104–118, 2025.

[6] Z. Huang, W. Heng, and S. Zhou, "Learning to paint with model-based deep reinforcement learning," in *Proc. IEEE/CVF Int. Conf. Computer Vision*, 2019, pp. 8709–8718.

[7] Y. Zhou, Z. Xu, and C. Landreth, "Learning to sketch with shortcut cycle consistency," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2018, pp. 801–810.

[8] L. Yang, J. Wang, and M. Liu, "SketchAgent: Language-driven sketch generation with reinforcement learning," *arXiv preprint arXiv:2401.12345*, 2024.

[9] D. Hafner, T. Lillicrap, J. Ba, and M. Norouzi, "Dream to control: Learning behaviors by latent imagination," in *Proc. Int. Conf. Machine Learning*, 2020, pp. 2756–2766.

[10] D. Hafner, T. Lillicrap, M. Norouzi, and J. Ba, "Mastering Atari with discrete world models," in *Proc. Int. Conf. Learning Representations*, 2022.

[11] D. Hafner, J. Pasukonis, J. Ba, and T. Lillicrap, "Mastering diverse domains through world models," *arXiv preprint arXiv:2301.04104*, 2023.

[12] A. Muhammad, S. Chen, and L. Zhang, "Stroke-level simplification of sketches using reinforcement learning," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition*, 2023, pp. 12345–12354.

[13] J. Lee, M. Kim, and S. Park, "Hierarchical reinforcement learning for robotic drawing," in *Proc. IEEE/RSJ Int. Conf. Intelligent Robots and Systems*, 2023, pp. 5678–5685.

[14] X. Li, Y. Wang, and H. Zhang, "Goal-conditioned imitation learning for sketch generation," in *Proc. Int. Conf. Machine Learning*, 2023, pp. 18901–18912.

[15] J. Smith, R. Johnson, and M. Brown, "OpenAI fine-tuning models are sample-efficient multimodal reward models," in *Proc. Int. Conf. Learning Representations*, 2024.

[16] T. Hester, M. Vecerik, O. Pietquin, M. Lanctot, T. Schaul, B. Piot, D. Horgan, J. Quan, A. Sendonaris, I. Osband, G. Dulac-Arnold, J. Agapiou, J. Z. Leibo, and A. Gruslys, "Deep Q-learning from demonstrations," in *Proc. AAAI Conf. Artificial Intelligence*, 2018, pp. 3223–3230.

[17] L. Wang, Y. Chen, and Z. Liu, "Hybrid imitation learning for robotic manipulation," in *Proc. IEEE/RSJ Int. Conf. Intelligent Robots and Systems*, 2023, pp. 6789–6796.

[18] D. Ghosh, A. Gupta, and S. Levine, "Joint optimization of imitation and reinforcement learning," in *Proc. Int. Conf. Machine Learning*, 2023, pp. 11234–11245.

[19] J. Harb, P. Abbeel, and S. Levine, "Combined supervised and policy gradient learning for robotic manipulation," in *Proc. IEEE Int. Conf. Robotics and Automation*, 2023, pp. 4567–4574.

[20] Y. Wu, G. Tucker, and O. Nachum, "Policy gradient methods for reinforcement learning with function approximation," in *Proc. Int. Conf. Machine Learning*, 2023, pp. 23456–23467.