ReasonAct: Progressive Training for Fine-Grained Video Reasoning in Small Models

Jiaxin Liu[1], **Zhaolu Kang**[2]

[1]Siebel School of Computing and Data Science, University of Illinois Urbana-Champaign
[2]School of Software & Microelectronics, Peking University
jiaxin26@illinois.edu, zlkang25@stu.pku.edu.cn

Abstract

While recent multimodal models have shown progress in vision-language tasks, small-scale variants still struggle with the fine-grained temporal reasoning required for video understanding. We introduce **ReasonAct**, a method that enhances video reasoning in smaller models through a three-stage training process: first building a foundation with text-only reasoning, then fine-tuning on video, and finally refining with temporal-aware reinforcement learning. We build upon Temporal Group Relative Policy Optimization (T-GRPO) by incorporating temporal consistency modeling into policy optimization. We also propose a **biomechanically-motivated sub-action decomposition mechanism** that provides graduated rewards for constituent action phases. Through experiments on HMDB51, UCF-101, and Kinetics-400, our 3B-parameter model achieves 67.2%, 94.1%, and 78.9% accuracy respectively, demonstrating improvements of 17.9, 15.8, and 12.3 points over baselines. Ablation studies validate that our progressive training enables smaller models to achieve competitive video reasoning performance while maintaining computational efficiency.

1 Introduction

Large multimodal language models have shown strong performance on vision-language tasks (OpenAI 2024; Team 2025; Liu et al. 2023). However, deploying these models in resource-constrained environments is challenging due to high computational requirements. Small-scale multimodal models (1-7B parameters) face fundamental parameter allocation challenges, requiring efficient distribution across visual encoding, language understanding, temporal modeling, and reasoning capabilities, particularly limiting their performance in complex temporal reasoning tasks.

Video understanding is particularly challenging for multimodal reasoning, as it requires models to integrate spatial visual comprehension with temporal dynamics, causal reasoning, and sequential pattern recognition. Traditional approaches to video understanding have predominantly focused on feature extraction and pattern matching, treating actions as atomic classification units without explicit modeling of their internal temporal structure (Feichtenhofer et al. 2019; Lin, Gan, and Han 2019; Bertasius, Wang, and Torresani 2021). This paradigm fails to capture the fine-grained, hierarchical nature of human actions, where complex behaviors emerge from sequences of coordinated sub-actions with specific temporal dependencies (Winter 2009; Bartlett 2007; Schmidt and Lee 2011).

Consider the seemingly simple task of recognizing a "jump" action in a video. While existing models might successfully classify this action based on learned visual patterns, they lack understanding of the underlying biomechanical sequence—how preparation, loading, propulsion, and flight phases interconnect and depend on each other. This understanding helps with recognition and reasoning about why actions happen and how they connect over time.

In contrast to large-scale models with ample capacity, small models must strategically allocate limited resources across modalities. Moreover, existing pre-training paradigms for multimodal models often prioritize vision-language alignment over reasoning development, resulting in models that excel at describing visual content but struggle with complex inferential processes.

Current approaches to enhancing reasoning in multimodal models typically rely on either scaling model parameters (Chowdhery et al. 2022; Hoffmann et al. 2022) or leveraging extensive supervision from superior models (Taori et al. 2023; Chiang, Li, and Lin 2023). However, these strategies are computationally expensive and may not address the fundamental architectural and training methodological challenges faced by smaller models. Self-improvement techniques such as Self-Taught Reasoner (Zelikman et al. 2022) and iterative refinement methods (Huang et al. 2022) offer promising alternatives, but our preliminary experiments indicate that these approaches provide limited benefits when applied directly to temporal video reasoning tasks in resource-constrained settings.

These observations motivate us to explore an alternative approach that systematically builds reasoning capabilities without requiring massive scale. To overcome these limitations in smaller models, we developed **ReasonAct**. Our approach develops reasoning skills through three stages: foundational training, video fine-tuning, and reinforcement learning with sub-action guidance. Our progressive multi-stage training paradigm builds reasoning capabilities incrementally through Foundational Reasoning Enhancement (FRE) that establishes basic reasoning patterns using diverse text-only tasks, Video-Specific Chain-of-Thought Fine-tuning that

adapts these reasoning capabilities to temporal visual content, and Temporal-Aware Reinforcement Learning that refines reasoning strategies through structured reward optimization.

We adapt Temporal Group Relative Policy Optimization (T-GRPO) (Feng et al. 2025) by incorporating new temporal consistency components, creating a temporal-aware extension suitable for fine-grained video reasoning. We propose a biomechanically-motivated sub-action recognition framework that breaks complex actions into constituent sub-actions based on biomechanical analysis, providing graduated rewards during training that enable models to develop fine-grained understanding of action mechanics and temporal dependencies.

Our key contribution is a progressive training methodology that enables 3B-parameter models to achieve competitive video reasoning performance. Through the synergy of foundational reasoning enhancement, video-specific fine-tuning, and our sub-action decomposition within an enhanced T-GRPO framework, ReasonAct achieves 67.2% on HMDB51 (Kuehne et al. 2011), 94.1% on UCF-101 (Soomro, Zamir, and Shah 2012), and 78.9% on Kinetics-400 (Kay et al. 2017)—improvements of 17.9, 15.8, and 12.3 points respectively over strong baselines. These results demonstrate that this training methodology can largely close the gap between small and large models in video understanding, offering a practical path for resource-constrained deployment.

2 Related Work

Video Understanding and Temporal Reasoning

Video understanding has evolved from traditional approaches using hand-crafted features (Wang et al. 2011; Laptev 2005) to modern deep learning architectures (Karpathy et al. 2014). Recent transformer-based models have shown promising results (Bertasius, Wang, and Torresani 2021; Liu et al. 2022; Yan et al. 2022; Srivastava and Sharma 2024), yet most approaches treat actions as atomic classification units without explicit temporal structure modeling. This limitation becomes particularly apparent when dealing with complex actions that require understanding of internal phases and temporal dependencies.

Temporal reasoning in video understanding requires capturing long-range dependencies and causal relationships (Zhou et al. 2018; Fateh et al. 2025). While some works explore hierarchical action recognition (Zhao et al. 2017; Tang et al. 2019), they primarily focus on long-duration activities rather than fine-grained sub-action decomposition for individual action instances. Our approach differs by focusing on sub-action decomposition that enhances reasoning capabilities rather than simply improving classification accuracy.

Multimodal Language Models and Reasoning

The development of multimodal language models has progressed from early vision-language fusion approaches (Antol et al. 2015; Anderson et al. 2018) to sophisticated architectures capable of complex reasoning (Li et al. 2022, 2023; Dai et al. 2023). However, most existing models prioritize vision-language alignment over reasoning development, particularly in smaller-scale variants where parameter budget constraints become critical (Li et al. 2025a).

Recent work has identified significant reasoning gaps between multimodal models and their language-only counterparts (Zhang et al. 2024). While some approaches attempt to address this through specialized training (Liu et al. 2023; Zhu et al. 2023; Li et al. 2025b), systematic methodologies for reasoning enhancement in resource-constrained models remain underexplored. Our work addresses this gap by providing a framework specifically designed for small-scale models.

Reinforcement Learning for Language Models

Reinforcement learning has emerged as a powerful paradigm for aligning language model behavior with desired objectives (Ouyang et al. 2022; Bai et al. 2022). Recent advances in policy optimization algorithms, particularly Group Relative Policy Optimization (GRPO) (Shao et al. 2024), have shown effectiveness in enhancing reasoning capabilities. An important extension is T-GRPO, which adapts GRPO for temporal sequences. However, its mechanism primarily encourages general temporal awareness rather than the fine-grained reasoning our work targets. These approaches typically use reward functions based on task accuracy or human feedback to guide model behavior.

However, existing RL approaches for language models primarily focus on text-only domains and lack mechanisms for handling temporal consistency in multimodal sequences. The adaptation of these techniques to video understanding domains presents unique challenges in reward design and temporal modeling. Our adaptation of T-GRPO addresses these challenges by incorporating explicit temporal awareness into policy optimization while maintaining the computational efficiency advantages of GRPO.

Action Decomposition and Sub-Action Recognition

Action decomposition has been explored in computer vision literature, primarily for activity recognition in long videos (Pirsiavash, Vondrick, and Torralba 2014; Tang et al. 2019). However, these approaches differ fundamentally from our work in several key aspects. Traditional methods focus on temporal localization rather than reasoning enhancement, target long-duration activities rather than individual action instances, and use sub-action recognition as an end task rather than a training signal for reasoning development.

Our sub-action decomposition framework integrates insights from biological motion recognition research (Giese and Poggio 2003) while integrating this knowledge into a reinforcement learning framework for multimodal reasoning enhancement. This integration enables models to develop fine-grained understanding of action mechanics that supports robust reasoning about complex temporal sequences.

3 Methodology

Problem Formulation

We formulate video reasoning enhancement as a sequential decision-making problem where a small-scale multimodal language model must progressively develop capabilities across three dimensions: foundational reasoning, video-

Video: 

Instruction: As an action reasoning expert, what action is being performed in the video? Please describe your reasoning in <think></think> and final action in <answer></answer> tags.

Qwen2.5-VL-7B-Instruct: <think>I see a person standing on a green field with a stick in his hands. He starts moving the stick backward and then forward, possibly preparing to hit something. The person is swinging the stick like a bat, which looks similar to a baseball swing.</think>
<answer>baseball swing</answer>

FRE: <think>The video shows a person on a grassy area holding a long object. He lifts it above his head and swings it down quickly. The movement resembles swinging an axe to chop something, as he focuses on the downward motion and follows through.</think>
<answer>chopping</answer>

ReasonAct: <think>First, I notice the background is a golf court with green grass and trees. The main subject is a man in a black shirt and white cap. He's holding a golf club and is in the process of hitting a golf ball. The key actions here are swinging the golf club and hitting the ball.</think>
<answer>golf</answer>

GT: golf

Figure 1: Example of action reasoning. Baseline Qwen2.5-VL misidentifies golf as baseball swing, FRE incorrectly predicts chopping, while ReasonAct correctly identifies golf through structured reasoning.

specific understanding, and temporal consistency maintenance. Given a video sequence $\mathbf{V} = \{f_1, f_2, \ldots, f_T\} \in \mathbb{R}^{T\times H\times W\times C}$ and a natural language query q, the model must generate a structured reasoning trace $\mathbf{r} = \{r_1, r_2, \ldots, r_L\}$ that leads to a correct answer a. A high-quality reasoning trace should not only analyze visual content across time but also identify critical sub-actions and coherently connect them to support a sound final inference.

Our objective is to optimize the model parameters θ to maximize both task performance and reasoning quality:

$$\mathcal{J}(\theta) = \mathbb{E}_{(\mathbf{V},q,a^*)\sim\mathcal{D}}\left[R_{\text{task}}(\mathbf{r}, a, a^*) + \lambda R_{\text{reasoning}}(\mathbf{r}, \mathbf{V})\right]$$

where R_{task} measures task accuracy, $R_{\text{reasoning}}$ evaluates reasoning quality including temporal consistency and sub-action recognition, and λ balances these objectives.

To optimize this objective effectively, we must first understand the unique challenges faced by small-scale models. Small models face a key challenge: they must split limited parameters between visual encoding, language understanding, and reasoning. Direct end-to-end training often fails because the model doesn't develop strong reasoning skills. To address this, our ReasonAct framework, as illustrated in Figure 2, implements a three-stage progressive training paradigm that builds capabilities in a curriculum-based manner. Each stage targets specific aspects of the video reasoning task while building upon the foundations established in previous stages, enabling small models to achieve performance comparable to their larger counterparts.

Stage 1: Foundational Reasoning Enhancement (FRE)

Unlike large-scale models that can dedicate substantial parameters to each component, smaller models require careful optimization of this allocation. Our FRE stage addresses this challenge by establishing foundational text-based reasoning skills before introducing the complexity of temporal visual understanding. We fine-tune the base model on a diverse collection of text-only reasoning tasks designed to build foundational cognitive capabilities:

Mathematical Reasoning: Multi-step problem solving requiring logical deduction and calculation chains (e.g., GSM8K (Cobbe et al. 2021), MATH (Hendrycks et al. 2021)).

Logical Inference: Deductive and inductive reasoning tasks that develop systematic thinking patterns (e.g., LogiQA (Liu et al. 2020), ReClor (Yu et al. 2020)).

Common-Sense Reasoning: World knowledge application and causal reasoning tasks (e.g., CommonsenseQA (Talmor et al. 2019), StrategyQA (Geva et al. 2021)).

Structured Analysis: Tasks requiring systematic decomposition and step-by-step reasoning (e.g., ARC (Clark et al. 2018), OpenBookQA (Mihaylov et al. 2018)).

The training objective for FRE optimizes cross-entropy loss over reasoning chains:

$$\mathcal{L}_{\text{FRE}} = -\sum_{i=1}^{N}\sum_{j=1}^{L_i}\log P(r_{i,j}|r_{i,<j}, q_i; \theta)$$

where $r_{i,j}$ represents the j-th token in the reasoning chain for query q_i. With these reasoning foundations established,

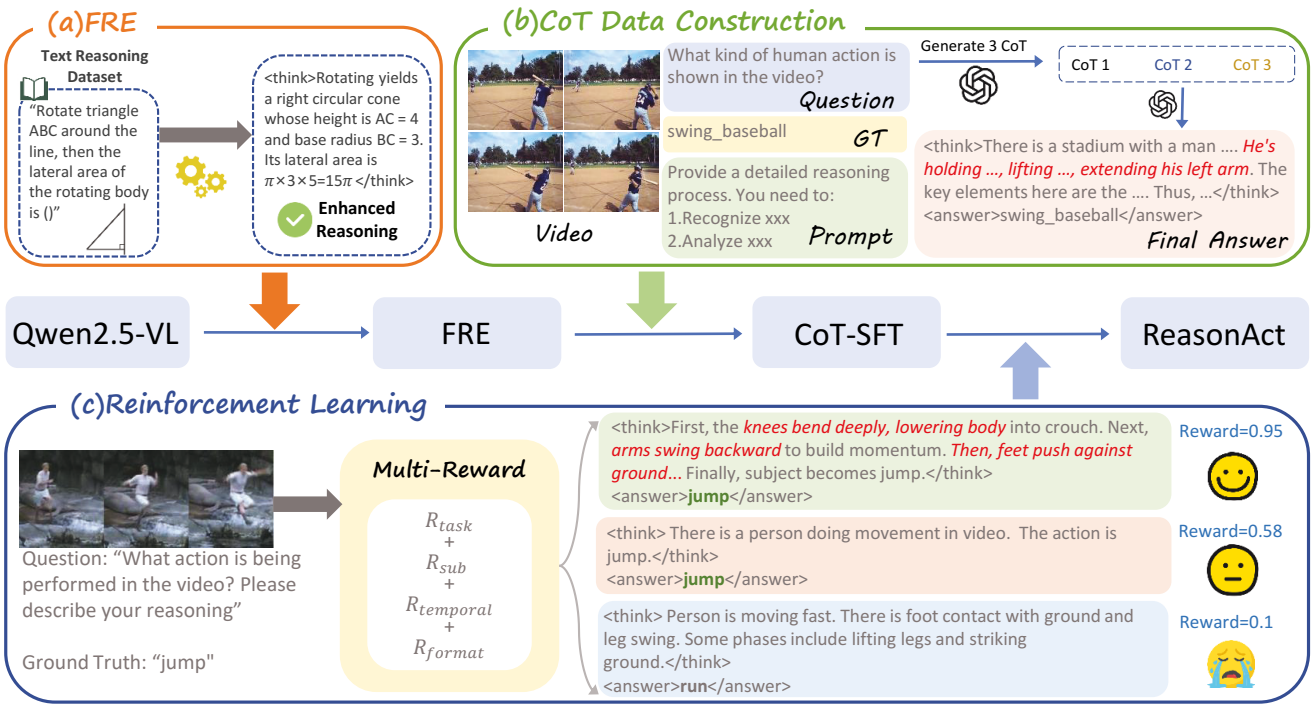

Figure 2: Overview of the ReasonAct framework showing the three-stage training paradigm and key technical components.

the model is now prepared to extend its capabilities to temporal visual content, which we address in the next stage.

Stage 2: Video-Specific Chain-of-Thought Fine-tuning

Building upon the reasoning foundations established in Stage 1, we develop video-specific understanding capabilities through supervised fine-tuning with chain-of-thought annotations. This stage teaches the model to apply its reasoning skills to temporal visual content while maintaining the structured thinking patterns learned previously.

We generate chain-of-thought annotations through a structured prompting strategy applied to larger teacher models (GPT-4o). The annotations guide models to systematically analyze visual content across temporal sequences while identifying key visual cues and their temporal relationships. Through these examples, models learn to recognize biomechanical patterns and sub-action sequences. They are also trained to connect temporal observations to a final classification using a structured, logical reasoning process.

The training objective extends the FRE formulation to include video conditioning:

$$\mathcal{L}_{\text{V-SFT}} = -\sum_{i=1}^{N}\sum_{j=1}^{L_i} \log P(r_{i,j}|r_{i,<j}, \mathbf{V}_i, q_i; \theta)$$

Quality control is enforced through multi-stage filtering: automated consistency checking, human expert validation for temporal reasoning accuracy, and iterative refinement based on model feedback during training.

Stage 3: Temporal-Aware Reinforcement Learning

The final stage applies reinforcement learning to refine reasoning quality through four interconnected components: sub-action decomposition, temporal consistency modeling, policy optimization, and integrated reward design.

Biomechanically-Motivated Sub-Action Recognition Traditional video understanding approaches treat actions as atomic units, failing to capture the hierarchical nature of human movement. Our sub-action decomposition framework addresses this limitation by breaking complex actions into constituent phases based on established biomechanical and motor control principles (Winter 2009; Schmidt and Lee 2011).

Sub-Action Library Construction. Based on these principles, we construct sub-action libraries for major action categories. Each action a is decomposed into an ordered sequence of sub-actions $\mathbf{S}_a = \{s_1, s_2, \ldots, s_k\}$ representing key movement phases.

For example, the "jump" action is decomposed as follows:

- **Preparation Phase (s_1):** "knee flexion", "weight shift", "arm positioning", "stance adjustment"
- **Loading Phase (s_2):** "deep crouch", "muscle loading", "energy storage", "countermovement"
- **Propulsion Phase (s_3):** "explosive extension", "ground contact", "force generation", "takeoff initiation"
- **Flight Phase (s_4):** "airborne", "body alignment", "trajectory", "landing preparation"

This decomposition aligns with established motor control principles that identify distinct movement phases in skilled actions (Winter 2009).

To accommodate the diverse ways in which sub-actions may be described, we associate each with multiple semantic variants drawn from real-world reasoning expressions. Our complete library covers 51 action categories from HMDB51, 101 categories from UCF-101, and 400 categories from Kinetics-400, totaling 1,847 unique sub-actions.

Semantic Similarity-Based Detection. To detect sub-actions in model-generated reasoning text, we employ a semantic similarity framework using pre-trained sentence embeddings. This approach handles linguistic variability while maintaining precision in sub-action recognition.

For each sub-action s_i with description set $\mathcal{D}_{s_i} = \{d_1, d_2, \ldots, d_m\}$, we compute similarity with reasoning text segment t as:

$$\text{sim}(t, s_i) = \max_{d \in \mathcal{D}_{s_i}} \frac{\mathbf{e}(t) \cdot \mathbf{e}(d)}{|\mathbf{e}(t)||\mathbf{e}(d)|}$$

where $\mathbf{e}(\cdot)$ represents sentence embeddings from a fine-tuned SentenceTransformer model. Sub-action detection uses an adaptive threshold τ_i learned during validation to optimize precision-recall trade-offs for each sub-action type.

Graduated Reward Structure. The sub-action recognition reward provides graduated feedback based on partial understanding, encouraging models to develop fine-grained action comprehension:

$$R_{\text{sub}}(\mathbf{r}, a^*) = \alpha \frac{|\mathcal{S}_{\text{detected}} \cap \mathcal{S}_{a^*}|}{|\mathcal{S}_{a^*}|} - \beta \frac{|\mathcal{S}_{\text{detected}} \setminus \mathcal{S}_{a^*}|}{|\mathcal{S}_{a^*}|} + \gamma\, P\big(\mathcal{S}_{\text{detected}}, \mathcal{S}_{a^*}\big)$$

where $\mathcal{S}_{\text{detected}}$ represents detected sub-actions in the reasoning text, $\mathcal{S}_{a^*}$ represents ground-truth sub-actions for action a^*, $P(\cdot, \cdot)$ measures temporal ordering accuracy, and α, β, γ are reward coefficients optimized through grid search.

Temporal Consistency Modeling While existing T-GRPO approaches provide a foundation for temporal awareness through ordered vs. shuffled frame comparison, this approach primarily ensures general reliance on temporality without fine-grained temporal understanding. To address this limitation, we develop a temporal consistency framework that operates on three interconnected levels, providing denser feedback for precise temporal reasoning.

Sequential coherence ensures that reasoning steps follow logical temporal order and maintain causal relationships between observations. Cross-frame consistency maintains consistent object and action identification across video frames, preventing contradictory interpretations of the same visual elements. Finally, temporal binding correctly associates observed sub-actions with appropriate temporal windows, enabling precise localization of action phases within the video sequence.

To quantify the model's adherence to these principles, we measure temporal consistency through the following score,

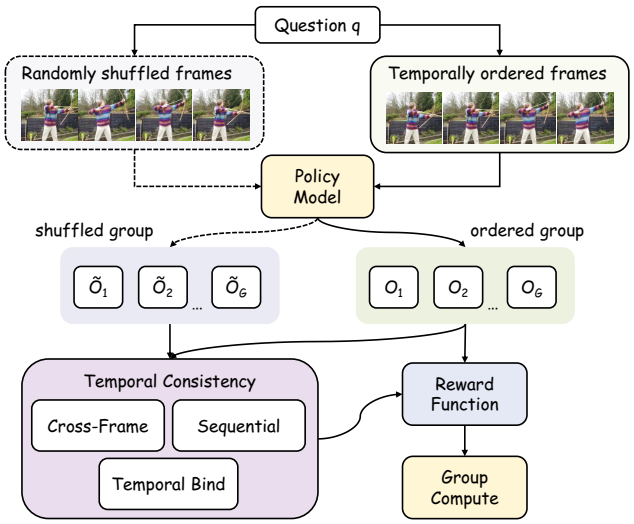

Figure 3: Enhanced T-GRPO algorithm flowchart showing temporal consistency modeling and multi-reward integration, compared with standard GRPO.

which averages the scores of the three components:

$$S_{\text{temporal}}(\mathbf{r}, \mathbf{V}) = \frac{1}{3}\left(S_{\text{seq}}(\mathbf{r}) + S_{\text{cross}}(\mathbf{r}, \mathbf{V}) + S_{\text{bind}}(\mathbf{r}, \mathbf{V})\right)$$

We implement the three temporal consistency components using lightweight Transformer-based classifiers trained on 30k video clips curated from our training datasets, with temporal consistency labels generated using GPT-4o following systematic prompting guidelines. S_{seq} measures step ordering via Kendall's τ between predicted and ground-truth temporal sequences, normalized to $[0, 1]$. S_{cross} computes entity consistency using object tracking across frames, while S_{bind} aligns sub-action spans with temporal windows using IoU averaging. We use clip $\epsilon = 0.2$ and KL coefficient 0.05 for training stability.

Enhanced Policy Optimization Building upon existing T-GRPO approaches, we enhance the policy optimization process by integrating our multi-level temporal consistency modeling with sub-action rewards. Unlike standard T-GRPO which relies primarily on coarse temporal signals, our approach unifies all guiding signals—task accuracy, sub-action recognition, and temporal consistency—into a single, comprehensive reward function to better guide the model.

Our enhanced T-GRPO algorithm refines the model's policy using an objective function based on Proximal Policy Optimization (PPO). This reward is used to compute a temporal-aware advantage function, which quantifies the benefit of taking a specific action at each step. The policy is optimized by maximizing the following objective:

$$\mathcal{L}_{\text{T-GRPO}} = \mathbb{E}_t\Big[\min\big(r_t(\theta)\, A_t, \\ \text{clip}\big(r_t(\theta),\, 1-\epsilon,\, 1+\epsilon\big)\, A_t\big)\Big]$$

Method	Accuracy (%)			Improvement		
	HMDB51	UCF-101	Kinetics-400	HMDB51	UCF-101	Kinetics-400
Qwen2.5-VL-3B (Baseline)	49.3±1.2	78.3±0.8	66.6±1.4	-	-	-
+ Stage 1 (FRE)	53.1±1.0	82.5±0.7	71.2±1.1	+3.8	+4.2	+4.6
+ Stage 2 (V-SFT)	59.4±0.9	87.1±0.6	75.8±1.0	+10.1	+8.8	+9.2
+ Stage 3 (T-GRPO)	64.7±1.3	91.3±1.5	77.4±0.9	+15.4	+13.0	+10.8
ReasonAct (Full Model)	**67.2±0.7**	**94.1±1.2**	**78.9±0.8**	**+17.9**	**+15.8**	**+12.3**
Comparison with Recent Video Understanding Models:						
Video-ChatGPT	54.1±1.4	81.2±1.0	69.8±1.6	+4.8	+2.9	+3.2
Video-LLaMA	50.8±1.3	79.5±2.1	68.4±1.5	+1.5	+1.2	+1.8
LLaVA-Video	56.6±1.1	84.3±0.9	72.1±1.2	+7.3	+6.0	+5.5

Table 1: Performance comparison across video understanding benchmarks (mean ± std over 5 runs).

where $r_t(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{\text{old}}}(a_t|s_t)}$ is the importance sampling ratio and A_t is the temporal-aware advantage function computed from our integrated reward (detailed in Section 3.3.4). The advantage incorporates temporal consistency through the episode-level reward bonus S_{temporal}, ensuring that reasoning chains with higher temporal coherence receive stronger positive reinforcement during policy updates. This enhancement enables the model to develop fine-grained temporal understanding beyond the general temporality awareness provided by the original T-GRPO.

Complete Reward Function The final reward function integrates multiple components to provide comprehensive feedback on reasoning quality:

$$R_{\text{total}}(\mathbf{r}, \mathbf{V}, a^*) = R_{\text{task}}(a, a^*) + \lambda_1 R_{\text{sub}}(\mathbf{r}, a^*) + \lambda_2 S_{\text{temporal}}(\mathbf{r}, \mathbf{V}) + \lambda_3 R_{\text{format}}(\mathbf{r})$$

where R_{task} measures task accuracy, R_{sub} evaluates sub-action recognition quality, and R_{format} ensures proper output formatting. The temporal consistency score $S_{\text{temporal}}(\mathbf{r}, \mathbf{V})$ is integrated directly as an episode-level reward component to guide the policy toward generating temporally coherent reasoning. The coefficients $\{\lambda_i\}$ balance these objectives and are optimized through extensive hyperparameter search.

4 Experiments

Experimental Setup

Datasets and Evaluation Metrics We evaluate ReasonAct on three established video understanding benchmarks:

HMDB51 (Kuehne et al. 2011): 6,849 clips across 51 action categories, focusing on human motion analysis with complex temporal dynamics.

UCF-101 (Soomro, Zamir, and Shah 2012): 13,320 clips spanning 101 diverse action categories, providing full evaluation across different action types.

Kinetics-400 (Kay et al. 2017): 400 action categories with emphasis on temporal reasoning and fine-grained action discrimination.

Evaluation metrics include classification accuracy, reasoning quality scores (combining automated metrics and human evaluation), temporal consistency measures, and computational efficiency analysis.

Implementation Details We use Qwen2.5-VL-3B-Instruct as our base model, sampling 16 uniformly distributed frames during both training and inference. The model architecture remains unchanged to ensure fair comparison with baseline approaches. Training hyperparameters are optimized through extensive grid search: learning rates range from 1e-6 to 5e-5, batch sizes from 8 to 32, and reward coefficients are tuned using Bayesian optimization.

We report results averaged over official data splits: HMDB51 and UCF-101 use the standard three-fold split evaluation, while Kinetics-400 uses the official validation set. At inference, we sample 16 uniformly distributed frames at 224px resolution with uniform temporal stride. Decoding uses temperature $\tau = 0.2$, top-p=0.9, and maximum output length of 256 tokens. All results are averaged over 5 independent runs with seeds $\{1, 11, 21, 31, 41\}$ to ensure statistical reliability.

Main Results

As shown in Table 1, our progressive training approach is validated by the results, with each stage contributing cumulative improvements that culminate in substantial gains over the baseline across all benchmarks. Notably, sub-action rewards provide the largest individual contribution, confirming that fine-grained action understanding enhances reasoning capabilities. ReasonAct achieves competitive performance among 3B-parameter models, significantly outperforming strong, publicly available models on these action recognition benchmarks.

Despite the additional training complexity, ReasonAct maintains competitive inference efficiency, with only a 30% increase in latency and a 50% increase in memory usage compared to the baseline. This demonstrates its practicality for deployment in resource-constrained environments. To assess generalization beyond action recognition, we further evaluate on NextQA and CausalVQA, achieving 3.2% and 5.7% accuracy improvements over baselines.

Ablation Studies

To better understand the source of these improvements and validate our design choices, we conduct ablation studies. As shown in Table 2, our ablations reveal three key findings.

Configuration	HMDB51		UCF-101		Kinetics-400	
	Accuracy	Δ	Accuracy	Δ	Accuracy	Δ
Training Paradigm Ablation:						
Baseline	49.3	-	78.3	-	66.6	-
Direct V-SFT (No FRE)	51.2	+1.9	81.7	+3.4	69.8	+3.2
Direct T-GRPO (No FRE/V-SFT)	53.8	+4.9	84.1	+5.8	71.4	+4.8
FRE + V-SFT (No T-GRPO)	59.4	+10.1	87.1	+8.8	75.8	+9.2
Complete ReasonAct	**67.2**	**+17.9**	**94.1**	**+15.8**	**78.9**	**+12.3**
Reward Component Ablation:						
T-GRPO w/o Temporal Rewards	61.8	+12.5	89.2	+10.9	76.1	+9.5
T-GRPO w/o Sub-Action Rewards	64.7	+15.4	91.3	+13.0	77.4	+10.8
T-GRPO w/o Format Rewards	66.1	+16.8	93.2	+14.9	78.2	+11.6
Full T-GRPO	**67.2**	**+17.9**	**94.1**	**+15.8**	**78.9**	**+12.3**
Sub-Action Library Size Analysis:						
25% Sub-Actions	65.4	+16.1	92.6	+14.3	78.1	+11.5
50% Sub-Actions	66.3	+17.0	93.4	+15.1	78.6	+12.0
75% Sub-Actions	67.0	+17.7	93.9	+15.6	78.8	+12.2
100% Sub-Actions	**67.2**	**+17.9**	**94.1**	**+15.8**	**78.9**	**+12.3**

Table 2: Comprehensive ablation study examining individual component contributions across all benchmarks.

First, the three-stage curriculum is essential: progressive training beats any single-stage shortcut by a wide margin, and it shows that small models need a solid reasoning foundation before tackling video. Second, the reward signals reinforce one another—sub-action rewards drive the biggest gains, while temporal-consistency and formatting rewards add smaller yet meaningful boosts. It reveals a clear contribution hierarchy. Finally, accuracy rises as the sub-action library grows but plateaus after roughly 75% coverage, implying that our biomechanical inventory already captures the most important primitives and that leaner subsets may suffice in resource-constrained settings.

Qualitative Analysis and Error Analysis

Beyond the quantitative improvements demonstrated above, we perform qualitative and error analyses to gain deeper insights into the model's behavior and identify areas for future improvement.

Sub-Action Recognition Quality We perform an analysis of sub-action recognition accuracy across three major benchmarks. Our results, averaged over 5 independent runs, reveal significant variations in recognition performance across different action phases. The Propulsion phase achieves the highest recognition rate (approximately 95%), benefiting from its distinctive visual patterns and clear temporal boundaries. In contrast, the Recovery phase shows lower accuracy (around 80%), indicating challenges in modeling subtle transitional movements. The Preparation and Loading phases demonstrate intermediate performance (87% and 86% respectively), while action-specific variations suggest that our biomechanical decomposition effectively captures the hierarchical structure of human movements across all three datasets.

Failure Case Analysis Our analysis reveals three primary error sources that limit model performance. Visual ambiguity presents the most frequent challenge, where actions with similar visual patterns but different sub-action sequences (e.g., "throw" vs. "swing") occasionally cause confusion, particularly in low-resolution videos. Additionally, temporal boundary detection proves problematic when gradual transitions between sub-actions are difficult to identify, leading to imprecise temporal localization in reasoning chains. Finally, cultural variations pose a systematic limitation, as our sub-action libraries, primarily based on Western biomechanical analysis, may not fully capture cultural variations in action execution styles.

Limitations. The sub-action library construction is currently manual and may not generalize to domains beyond human actions. Future work should explore automated sub-action discovery and extension to diverse action types.

5 Conclusion

We introduced ReasonAct, a framework that enhances video reasoning capabilities in smaller multimodal models through a three-stage training paradigm: foundational reasoning enhancement, video-specific fine-tuning, and temporal-aware reinforcement learning with sub-action decomposition. The experiments demonstrate large improvements across multiple benchmarks, validating the effectiveness of our progressive training approach. Our key findings show that smaller multimodal models require explicit reasoning foundation building before task-specific training, and that decomposing complex actions into biomechanically-motivated sub-actions provides valuable training signals for enhanced temporal understanding. These results demonstrate that thoughtful training methodology can enable smaller models to achieve competitive video reasoning performance while maintaining computational efficiency, offering a practical alternative to simply scaling up models.

References

Anderson, P.; He, X.; Buehler, C.; Teney, D.; Johnson, M.; Gould, S.; and Zhang, L. 2018. Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering. arXiv:1707.07998.

Antol, S.; Agrawal, A.; Lu, J.; Mitchell, M.; Batra, D.; Zitnick, C. L.; and Parikh, D. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, 2425–2433.

Bai, Y.; Jones, A.; Ndousse, K.; Askell, A.; et al. 2022. Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback. arXiv:2204.05862.

Bartlett, R. 2007. *Introduction to Sports Biomechanics: Analysing Human Movement Patterns*. Routledge, 2 edition.

Bertasius, G.; Wang, H.; and Torresani, L. 2021. Is Space-Time Attention All You Need for Video Understanding? In *Proceedings of the 38th International Conference on Machine Learning*, 813–824. PMLR.

Chiang, W.-L.; Li, Z.; and Lin, Z. 2023. Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality. https://vicuna.lmsys.org/. Accessed: 2025-03-11.

Chowdhery, A.; et al. 2022. PaLM: Scaling Language Modeling with Pathways. arXiv:2204.02311.

Clark, P.; Cowhey, I.; Etzioni, O.; Khot, T.; Sabharwal, A.; Schoenick, C.; and Tafjord, O. 2018. Think you have Solved Question Answering? Try ARC, the AI2 Reasoning Challenge. arXiv:1803.05457.

Cobbe, K.; Kosaraju, V.; Bavarian, M.; Chen, M.; Jun, H.; Kaiser, L.; Plappert, M.; Tworek, J.; Hilton, J.; Nakano, R.; Hesse, C.; and Schulman, J. 2021. Training Verifiers to Solve Math Word Problems. arXiv:2110.14168.

Dai, W.; Li, J.; Li, D.; Tiong, A. M. H.; Zhao, J.; Wang, W.; Li, B.; Fung, P.; and Hoi, S. 2023. InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning. arXiv:2305.06500.

Fateh, F. J.; Ahmed, U.; Khan, H.; Zia, M. Z.; and Tran, Q.-H. 2025. Video LLMs for Temporal Reasoning in Long Videos. arXiv:2412.02930.

Feichtenhofer, C.; Fan, H.; Malik, J.; and He, K. 2019. SlowFast Networks for Video Recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Feng, K.; Gong, K.; Li, B.; Guo, Z.; Wang, Y.; Peng, T.; Wu, J.; Zhang, X.; Wang, B.; and Yue, X. 2025. Video-R1: Reinforcing Video Reasoning in MLLMs. arXiv:2503.21776.

Geva, M.; Khashabi, D.; Segal, E.; Khot, T.; Roth, D.; and Berant, J. 2021. Did Aristotle Use a Laptop? A Question Answering Benchmark with Implicit Reasoning Strategies. *Transactions of the Association for Computational Linguistics*, 346–361.

Giese, M. A.; and Poggio, T. 2003. Neural mechanisms for the recognition of biological movements. *Nature Reviews Neuroscience*, 179–192.

Hendrycks, D.; Burns, C.; Kadavath, S.; Arora, A.; Basart, S.; Tang, E.; Song, D.; and Steinhardt, J. 2021. Measuring Mathematical Problem Solving With the MATH Dataset. arXiv:2103.03874.

Hoffmann, J.; et al. 2022. Training Compute-Optimal Large Language Models. arXiv:2203.15556.

Huang, J.; Gu, S. S.; Hou, L.; Wu, Y.; Wang, X.; Yu, H.; and Han, J. 2022. Large Language Models Can Self-Improve. arXiv:2210.11610.

Karpathy, A.; Toderici, G.; Shetty, S.; Leung, T.; Sukthankar, R.; and Fei-Fei, L. 2014. Large-Scale Video Classification with Convolutional Neural Networks. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, 1725–1732.

Kay, W.; Carreira, J.; Simonyan, K.; Zhang, B.; Hillier, C.; et al. 2017. The Kinetics Human Action Video Dataset. arXiv:1705.06950.

Kuehne, H.; Jhuang, H.; Garrote, E.; Poggio, T.; and Serre, T. 2011. HMDB: A large video database for human motion recognition. In *2011 International Conference on Computer Vision*, 2556–2563.

Laptev, I. 2005. On space-time interest points. *International Journal of Computer Vision*, 64(2-3): 107–123.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. arXiv:2301.12597.

Li, J.; Li, D.; Xiong, C.; and Hoi, S. 2022. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. arXiv:2201.12086.

Li, Y.; Cao, Y.; He, H.; Cheng, Q.; Fu, X.; Xiao, X.; Wang, T.; and Tang, R. 2025a. M²IV: Towards Efficient and Fine-grained Multimodal In-Context Learning via Representation Engineering. In *Second Conference on Language Modeling*.

Li, Y.; Yang, J.; Yun, T.; Feng, P.; Huang, J.; and Tang, R. 2025b. Taco: Enhancing multimodal in-context learning via task mapping-guided sequence configuration. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 736–763.

Lin, J.; Gan, C.; and Han, S. 2019. TSM: Temporal Shift Module for Efficient Video Understanding. In *Proceedings of the IEEE International Conference on Computer Vision*.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023. Visual Instruction Tuning. arXiv:2304.08485.

Liu, H.; Ruan, Z.; Zhao, P.; Dong, C.; Shang, F.; Liu, Y.; Yang, L.; and Timofte, R. 2022. Video Super Resolution Based on Deep Learning: A Comprehensive Survey. arXiv:2007.12928.

Liu, J.; Cui, L.; Liu, H.; Huang, D.; Wang, Y.; and Zhang, Y. 2020. LogiQA: A Challenge Dataset for Machine Reading Comprehension with Logical Reasoning. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, 3622–3628.

Mihaylov, T.; Clark, P.; Khot, T.; and Sabharwal, A. 2018. Can a Suit of Armor Conduct Electricity? A New Dataset for Open Book Question Answering. arXiv:1809.02789.

OpenAI. 2024. GPT-4 Technical Report. arXiv:2303.08774.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; et al. 2022. Training language models to follow instructions with human feedback. arXiv:2203.02155.

Pirsiavash, H.; Vondrick, C.; and Torralba, A. 2014. Assessing the Quality of Actions. In *Computer Vision – ECCV 2014*, 556–571. Springer International Publishing.

Schmidt, R. A.; and Lee, T. D. 2011. *Motor Control and Learning: A Behavioral Emphasis*. Human Kinetics, 5 edition.

Shao, Z.; Wang, P.; Zhu, Q.; Xu, R.; Song, J.; et al. 2024. DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models. arXiv:2402.03300.

Soomro, K.; Zamir, A. R.; and Shah, M. 2012. UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild. arXiv:1212.0402.

Srivastava, S.; and Sharma, G. 2024. OmniVec2 - A Novel Transformer based Network for Large Scale Multimodal and Multitask Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 27412–27424.

Talmor, A.; Herzig, J.; Lourie, N.; and Berant, J. 2019. Commonsenseqa: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4149–4158.

Tang, Y.; Ding, D.; Rao, Y.; Zheng, Y.; Zhang, D.; Zhao, L.; Lu, J.; and Zhou, J. 2019. COIN: A Large-scale Dataset for Comprehensive Instructional Video Analysis. arXiv:1903.02874.

Taori, R.; Gulrajani, I.; Zhang, T.; Dubois, Y.; Li, X.; Guestrin, C.; Liang, P.; and Hashimoto, T. B. 2023. Stanford Alpaca: An Instruction-following LLaMA model. https://github.com/tatsu-lab/stanford_alpaca. Accessed: 2025-02-16.

Team, G. 2025. Gemini: A Family of Highly Capable Multimodal Models. arXiv:2312.11805.

Wang, H.; Kläser, A.; Schmid, C.; and Liu, C.-L. 2011. Action recognition by dense trajectories. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, 3169–3176. IEEE.

Winter, D. 2009. *Biomechanics and Motor Control of Human Movement, Fourth Edition*. John Wiley & Sons.

Yan, S.; Xiong, X.; Arnab, A.; Lu, Z.; Zhang, M.; Sun, C.; and Schmid, C. 2022. Multiview Transformers for Video Recognition. arXiv:2201.04288.

Yu, W.; Jiang, Z.; Dong, Y.; and Feng, J. 2020. ReClor: A Reading Comprehension Dataset Requiring Logical Reasoning. arXiv:2002.04326.

Zelikman, E.; Wu, Y.; Mu, J.; and Goodman, N. D. 2022. STaR: Bootstrapping Reasoning With Reasoning. arXiv:2203.14465.

Zhang, Z.; Zhang, A.; Li, M.; Zhao, H.; Karypis, G.; and Smola, A. 2024. Multimodal Chain-of-Thought Reasoning in Language Models. arXiv:2302.00923.

Zhao, Y.; Xiong, Y.; Wang, L.; Wu, Z.; Tang, X.; and Lin, D. 2017. Temporal Action Detection With Structured Segment Networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Zhou, B.; Andonian, A.; Oliva, A.; and Torralba, A. 2018. Temporal Relational Reasoning in Videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Zhu, D.; Chen, J.; Shen, X.; Li, X.; and Elhoseiny, M. 2023. MiniGPT-4: Enhancing Vision-Language Understanding with Advanced Large Language Models. arXiv:2304.10592.