Reason in the Words You Speak: Idiolectal Paraphrasing Off-Policy Traces for Reasoning Distillation in VideoLLMs

Ji Soo Lee[1] **Jinyoung Park**[1] **Seohyun Lee**[1] **Jongha Kim**[2] **Joonmyung Choi**[2]
Jinsung Yoon[3] **Hyunwoo J. Kim**[1*]

[1]KAIST, [2]Korea University, [3]Google Cloud AI Research

{jislee,jinyoung.park,seohyunlee,hyunwoojkim}@kaist.ac.kr
{jonghakim,pizard}@korea.ac.kr
jinsungyoon@google.com

Abstract

Recent large language models achieve strong performance on complex reasoning tasks, where reinforcement learning with Group Relative Policy Optimization (GRPO) has emerged as a leading paradigm for optimizing models on self-generated trajectories. However, the on-policy nature of GRPO bounds the model to the reasoning skills it can already produce, restricting to learn more advanced capabilities. Prior works inject privileged reasoning traces from a stronger teacher policy to guide training, yet these traces are inherently out of distribution with respect to the student policy. We observe that this mismatch between on-policy and off-policy causes gradient clipping on semantically critical reasoning tokens, ultimately rewarding correct answers while leaving the reasoning that justifies them unlearned. Hence, we propose **Echo-GRPO**, a framework that lets the model reason in the words it speaks. Rather than imitating low-probability privileged traces from the teacher model, Echo-GRPO rewrites them into the student policy's own *idiolect*, that is, its own characteristic vocabulary and expression patterns, while preserving their semantics via Dual-Reference Decoding. We instantiate this framework as **VideoEcho-R1** for video reasoning distillation, achieving consistent improvements across three multimodal LLM backbones and five benchmarks. Finally, we show that our idiolectal paraphrasing is a plug-in module that consistently improves both RL and supervised fine-tuning frameworks for reasoning distillation, demonstrating that policy-aligned supervision extends beyond GRPO.

1 Introduction

Recent advances in large language models have enabled significant progress on complex reasoning tasks, where Group Relative Policy Optimization (GRPO) [1] has emerged as an effective paradigm for reinforcing reasoning over self-generated trajectories [2–5]. Yet on-policy GRPO is inherently bound by its own policy formulation, from which the model can only learn from its existing reasoning skills. To address this limitation, prior work [6, 7] introduces Mixed-Policy GRPO, which injects privileged reasoning traces from a stronger teacher policy by replacing one of the sampled rollouts.

While this provides advanced guidance, it introduces a fundamental challenge. Privileged traces are typically out-of-distribution with respect to the current policy, sampled from a stronger teacher policy rather than the student policy itself, which poses a challenge for effective learning. This distributional

*Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

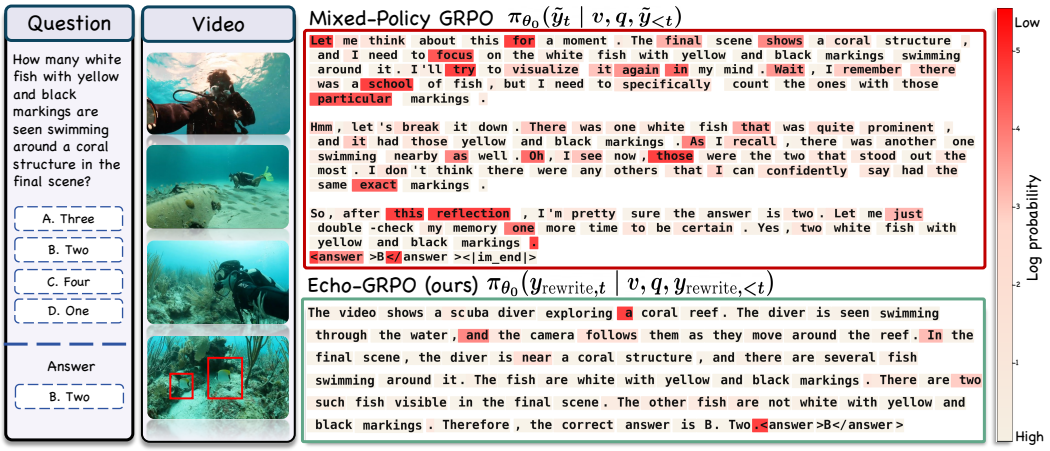


Figure 1: **Reasoning in the words you know.** Token-level log-probabilities under the policy π_θ (red shading indicates low probability). *Top:* Mixed-Policy GRPO forces the policy to imitate privileged reasoning traces from a stronger teacher, leaving semantically critical ones, such as "school of fish," in low-likelihood regions. *Bottom:* Echo-GRPO paraphrases the same reasoning into the policy's native distribution ("school of fish" $\rightarrow$ "several fish"), preserving the semantics while keeping all tokens within the policy's distribution.

gap is especially pronounced in VideoLLMs [8–12], where reasoning requires complex cross-modal interactions across visuals and text, further distancing the student policy from the teacher's trace distribution. Concretely, as shown in Fig. 1, most tokens in the privileged trace carry low likelihood under the policy's native distribution. Since the student policy is unlikely to generate those tokens, the importance sampling ratio of GRPO deviates from unity, triggering trust-region clipping that suppresses their gradient updates and excludes them from training entirely. Critically, we observe that those clipped tokens are not noise but semantically essential reasoning components, ultimately rewarding the model for the right answer while not learning the reasoning that justifies it.

To address this, we propose **Echo-GRPO**, a rewriting framework that makes the model reason in the words it speaks. Rather than imitating out-of-distribution traces, Echo-GRPO rewrites them with respect to the student's own *idiolect*, that is, its own characteristic vocabulary and expression patterns, while effectively preserving their semantics via Dual-Reference Decoding (DRD). DRD combines two conditional distributions of the initial policy in a product-of-experts formulation: a semantic reference that constrains token selection to be faithful to the privileged trace, and a distributional reference that ensures each token remains probable under the student policy. As illustrated in Fig. 1, the privileged trace describes a group of fish as a 'school of fish', while the rewritten trajectory expresses the same observation as 'several fish', which lies within the policy's distribution while preserving the semantic meaning.

We demonstrate that **VideoEcho-R1**, optimized with Echo-GRPO, achieves consistent improvements in reasoning distillation across three multimodal LLM backbones and five benchmarks, and our idiolectal rewriting further serves as a plug-in module that improves both reinforcement learning and supervised fine-tuning frameworks.

Our contributions can be summarized as:

- We identify the underexplored failure mode of mixed-policy GRPO, where the trust-region clipping suppresses gradient updates on semantically critical reasoning tokens, rewarding correct answers while leaving the underlying reasoning unlearned.
- We propose **Echo-GRPO**, an idiolectal paraphrasing framework that aligns off-policy privileged traces with the policy's native distribution via Dual-Reference Decoding that promotes paraphrased trajectories to be both semantically faithful and distributionally coherent.
- We introduce **VideoEcho-R1**, the model trained with Echo-GRPO for video reasoning distillation, which outperforms vanilla GRPO, supervised fine-tuning, and Mixed-Policy

GRPO across three multimodal LLM backbones (InternVL3.5-4B, Qwen3-VL-4B, Qwen3-VL-8B) and five video reasoning benchmarks.

- Idiolectal paraphrasing is a plug-in module that consistently improves reasoning distillation across both reinforcement learning frameworks and supervised fine-tuning, demonstrating that policy-aligned supervision extends well beyond vanilla GRPO.

2 Related Works

Reinforcement Learning in LLM Reasoning. Reinforcement learning [13–20] has emerged as a key driver for advancing LLM reasoning, demonstrated by recent systems such as OpenAI o1 [4], DeepSeek-R1 [2], and Kimi-1.5 [3]. GRPO [1, 2] has emerged as a widely adopted on-policy paradigm, with subsequent work refining it to address trust-region behavior, length bias, and KL regularization [5, 21–23]. A parallel line injects privileged supervision from stronger teachers to overcome the on-policy capability ceiling: Mixed-Policy GRPO [24, 6] substitutes one rollout with an off-policy trace, LUFFY [6] further applies regularized importance sampling to balance imitation and exploration, and OPSD [25] minimizes per-token divergence between a privileged-context teacher and the student over student-generated rollouts. Our work identifies a specific failure mode in this setting: gradient suppression from trust-region clipping on semantically critical tokens, and addresses it with idiolectal paraphrasing jointly faithful to the teacher's semantics and probable under the policy, directly stabilizing importance sampling ratios without requiring gradient or loss level correction.

Reasoning in Multimodal Large Language Models. Recent works extend RL-based reasoning to multimodal LLMs across diverse visual tasks [26–28, 7, 29–34]. Vision-R1 [26] adapts R1-style RL to image reasoning, while Video-R1 [35], VideoChat-R1 [28], and Tempsamp-R1 [7] extend it to video, targeting the temporal space of temporal-aware GRPO and rewards, respectively. Reason-RFT [36] further explores self-reflection and staged fine-tuning for vision-language reasoning. While these works successfully transplant RL-for-reasoning to the multimodal regime, the off-policy distillation gap is particularly acute in the video setting, where traces from large multimodal teachers lie far outside the distribution of smaller VideoLLMs, and semantic grounding in spatiotemporal evidence makes clipping-induced suppression especially damaging.

3 Method

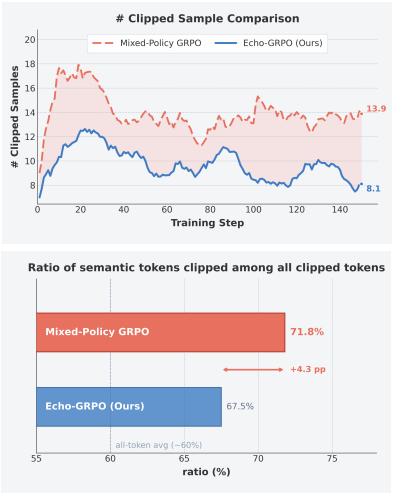
In this section, we first provide a brief overview of GRPO for reasoning distillation and the assumption it relies on (Sec. 3.1). Then we identify the failure mode of Mixed-Policy GRPO, where trust-region clipping suppresses gradient updates on semantically critical reasoning tokens (Sec. 3.2). To address this, we propose Echo-GRPO, an idiolectal paraphrasing framework that aligns the privileged reasoning traces into the policy's native distribution via Dual-Reference Decoding (Sec. 3.3).

3.1 Reasoning Distillation with GRPO

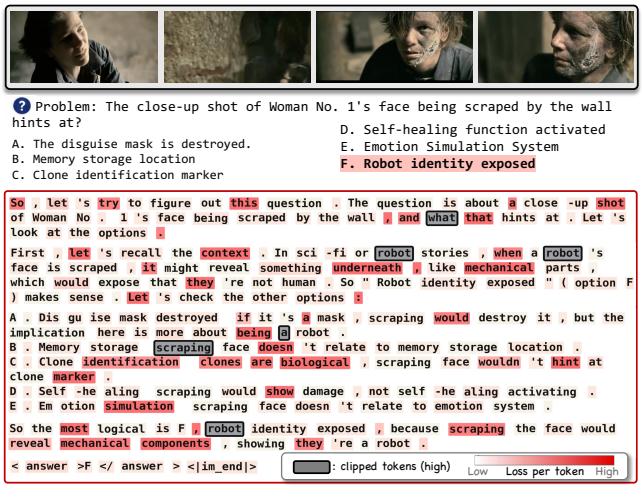
Group Relative Policy Optimization (GRPO) [1] serves as the underlying framework for reasoning distillation, where an off-policy reasoning trace is adopted as a supervision signal to guide optimization, also referred to as Mixed-Policy GRPO [6]. Given an input (v, q) consisting of a video and a question, we consider two types of reasoning traces: (1) *naive* reasoning traces $y \sim \pi_{\text{old}}(\cdot \mid v, q)$ sampled from the current policy, and (2) *privileged* reasoning trace $\tilde{y} \sim \pi_T(\cdot \mid v, q)$ obtained from a stronger teacher policy π_T. Then we combine the set of trajectories as $\mathcal{G} = \{y_i\}_{i=1}^{G-1} \cup \{\tilde{y}\}$ where G denotes the number of candidate samples. The GRPO optimizes the following objective over $\mathcal{G}$:

$$J_{\text{GRPO}}(\theta) = \mathbb{E}_{v,q\sim\mathcal{D},\{y_i\}_{i=1}^{G-1}\sim\pi_{\text{old}}(\cdot|v,q),\tilde{y}\sim\pi_T(\cdot|v,q)} \left[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|y_i|}\sum_{t=1}^{|y_i|}\min\left(\hat{r}_{i,t}(\theta)\hat{A}_i, \text{clip}\left(\hat{r}_{i,t}(\theta), 1-\epsilon, 1+\epsilon\right)\hat{A}_i\right)\right], \tag{1}$$

where ϵ is the clipping hyperparameter, and the importance sampling ratio is $\hat{r}_{i,t} = \frac{\pi_\theta(y_{i,t}|v,q,y_{i,<t})}{\pi_{\text{old}}(y_{i,t}|v,q,y_{i,<t})}$. The advantage $\hat{A}_i$ for reasoning trace y_i is computed using a group-normalized reward $\hat{A}_i =$



(a) **Clipping rate comparison.**



(b) **Example of Semantically important token clipped.**

Figure 2: (a) (top) Number of samples containing at least one clipped token throughout training for Mixed-Policy GRPO (red), and Echo-GRPO (blue). (bottom) Ratio of semantically important tokens (*e.g.*, nouns from the question and options) among all clipped tokens for each method. (b) Token-level visualization of clipping behavior on a representative video reasoning sample, where grey boxes indicate clipped tokens and red shading denotes per-token loss intensity.

$\frac{\mathcal{R}(y_i)-\mu_{\mathcal{G}}}{\sigma_{\mathcal{G}}}$, where μ and $\sigma_{\mathcal{G}}$ are the mean and standard deviation of a set of rewards from reasoning traces $\mathcal{G}$. Following prior works [5, 21], we omit the KL divergence penalty.

The importance sampling ratio $\hat{r}_{i,t}(\theta)$ plays a central role in ensuring valid policy updates when optimizing over trajectories sampled from a previous policy close to the current policy. Crucially, this mechanism relies on the assumption that all training trajectories in $\mathcal{G}$ lie within the behavior (old) policy distribution, keeping importance ratios near unity and gradient updates stable. However, adopting the privileged trace $\tilde{y}$ that is not sampled from π_θ but from a separate teacher policy π_T may break this assumption.

3.2 Clipping-Induced Suppression of Essential Reasoning

Substituting an off-policy privileged trace $\tilde{y}$ into the GRPO objective creates a cascading failure where the distributional gap between the π_θ and π_T first manifests as instability in the importance sampling ratio $\hat{r}_{i,t}(\theta)$. Specifically, the model is forced to update its policy towards a rigid, out-of-distribution reasoning trace, $\tilde{y}$, which it is highly unlikely to generate, *i.e.*, low $\pi_\theta(\tilde{y}|v, q)$.

Trust-region clipping discards a substantial fraction of samples. This instability propagates directly into trust-region clipping, where, when $\hat{r}_{i,t}(\theta)$ falls outside the trust region of $[1-\epsilon, 1+\epsilon]$, the corresponding gradient update is discarded. In practice, since privileged traces carry positive advantages while their tokens remain low-probability under the student policy, the ratio tends to exceed $1+\epsilon$, consistently triggering gradient suppression. In Fig. 2a (top), we measure the fraction of training samples containing at least one clipped token, where we find that Mixed-Policy GRPO clips a substantial fraction throughout optimization (dashed red).

Not all clipped tokens are noisy. To better understand the impact, we visualize token-level clipping in Fig. 2b. Tokens excluded from gradient updates are highlighted in grey, with per-token loss intensity shaded in red. As depicted, the entity "robot", the key referent for the answer "Robot identity exposed", is clipped from the gradient update, which indicates the trajectory received a high reward (the model arrives at the correct answer), but the gradient signal that would teach the model why the answer is correct is silently suppressed. Similarly, as shown in Fig. 2a (bottom), semantically important tokens (*e.g.*, nouns from the question and options) account for a disproportionately large share of clipped tokens throughout training, with 71.8% in Mixed-Policy GRPO.

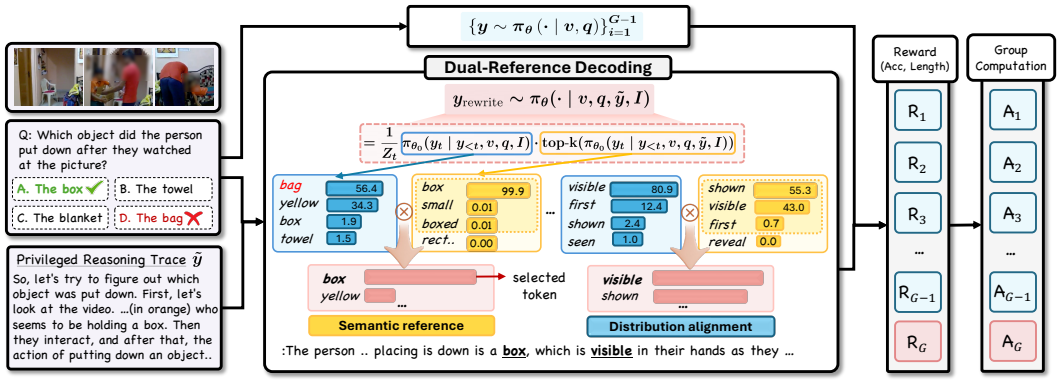


Figure 3: **Overview of Echo-GRPO.** Given a video question and a privileged reasoning trace $\tilde{y}$, Echo-GRPO adopts an idiolectal paraphrased trace y_{rewrite} produced by Dual-Reference Decoding (DRD). DRD combines two conditional distributions of π_{θ_0} as a product of distributions: the semantic reference (yellow) constrains token selection to be faithful to $\tilde{y}$, correcting the base policy's preference for semantically incorrect tokens (*e.g.*, 'bag' to 'box'); while distribution alignment (blue) resolves ties among semantically equivalent candidates by preferring tokens the student policy is more likely to generate (*e.g.*, 'visible' over 'shown'), keeping the rewritten trace aligned to its naive distribution.

Rewriting Prompt

```
Reference:
{{PRIVILEGED REASONING TRACE}}

Answer the question below based on the video.
Do not mention the reference.

Question:
{{QUESTION}}
```

Figure 4: **Prompt for idiolectal paraphrasing.** The privileged reasoning trace $\tilde{y}$ corresponds to (`{{PRIVILEGED REASONING TRACE}}`). We append the question and options at the end of the prompt. For the Question and Options we follow the format of OneThinker [27]. Further details are in the supplement.

To directly test whether clipping semantically important tokens contributes to this failure, we vary which tokens are unclipped while keeping all other settings fixed across three major benchmarks. We compare unclipping all tokens, a randomly selected subset, and a subset of semantically important tokens. As shown in Tab. 1, unclipping all tokens improves Mixed-Policy from 35.9 to 40.8, while unclipping a random subset reaches 48.4. Selectively unclipping semantic tokens yields the largest gain among these variants, reaching 52.0 (+16.1). Since the random and semantic variants unclip the same number of tokens and differ only in which tokens are selected, the higher performance of semantic unclipping supports the view that clipping semantically important tokens is a key factor behind Mixed-Policy failure.

Table 1: **Effect of clipping on Mixed-Policy failure.** OOD and ID refer to Out-of-distribution and In-distribution, respectively. Performance is averaged over three major benchmarks: VideoMMMU, MMVU, and Video-Holmes.

Clipping Method	Priv. Trace	Avg.	Δ
Mixed-Policy	OOD	35.9	–
+ Unclip all tokens	OOD	40.8	+4.9
+ Unclip random tokens	OOD	48.4	+12.5
+ Unclip semantic tokens	OOD	**52.0**	**+16.1**
Echo-GRPO (Ours)	ID	**58.2**	**+22.3**

3.3 Idiolectally Paraphrasing Privileged Traces

To address the aforementioned limitations, we propose **Echo-GRPO**, which resolves this by replacing the privileged trace $\tilde{y}$ with an idiolectally paraphrased trace y_{rewrite} that preserves the semantics of

$\tilde{y}$ while lying within the policy's native distribution, *i.e.*, reasoning in the phrases that the model speaks with. The goal of Echo-GRPO is to seek a rewritten trajectory y_{rewrite} that satisfies two properties: (1) preserve the semantic content of $\tilde{y}$ (semantic coherence), and (2) each token should lie within the policy's native distribution (distribution alignment). A simple approach is sampling $y_{\text{rewrite}} \sim \pi_\theta(\cdot|v, q, \tilde{y}, I)$ where I is a rewriting instruction prompt (shown in Fig 4). While simple rewriting conditioned on $\tilde{y}$ is effective (Sec. 4.4), it shifts the generation distribution toward the privileged trace, where we want to ensure that generated tokens lie within the policy's native distribution *without* conditioning on $\tilde{y}$.

Dual-Reference Decoding (DRD). To resolve this conditional shift, we introduce Dual-Reference Decoding (DRD), a product-of-experts formulation that combines two conditional distributions of π_{θ_0}, where π_{θ_0} is the initial policy model. The first serves as a semantic reference, conditioning on the privileged trace and contributing its top-k candidates, while the second acts as a distributional reference, evaluating the policy *without* privileged conditioning over the full vocabulary. At every step, the two distributions are multiplied so that the token receives high probability only when the token falls under both references. Formally, it can be written as:

$$
\begin{aligned}
\pi_{\text{rewrite}}\left(y_t \mid y_{<t}, v, q, \tilde{y}, I\right) &= \frac{1}{Z_t}\left[\underbrace{\pi_{\theta_0}\left(y_t \mid y_{<t}, v, q, I\right)}_{\text{Distributional reference}} \cdot \underbrace{\text{top-}k\left(\pi_{\theta_0}(y_t \mid y_{<t}, v, q, \tilde{y}, I)\right)}_{\text{Semantic reference}}\right], \\
\text{top-}k\big(\pi_{\theta_0}(y_t \mid y_{<t}, v, q, \tilde{y}, I)\big) &= \pi_{\theta_0}(y_t \mid y_{<t}, v, q, \tilde{y}, I) \cdot \mathbf{1}_{\{y_t \in \mathcal{V}_k^{(t)}\}}, \\
\mathcal{V}_k^{(t)} &= \underset{V \subset \mathcal{V}, |V|=k}{\arg\max} \sum_{y_t \in V} \pi_{\theta_0}(y_t \mid y_{<t}, v, q, \tilde{y}, I).
\end{aligned}
\tag{2}
$$

The set $\mathcal{V}_k^{(t)}$ is the set of top-k tokens under the semantic reference, and Z_t is a normalization constant. DRD actively shapes both distributions for the final probability for the current step token, of which the tokens that fall outside $\mathcal{V}_k^{(t)}$ will not be preferred as affected by the indicator $\mathbf{1}(\cdot)$. Hence, a token is more favored when it is plausible, which is consistent with the privileged reasoning trace's semantic meaning and carries meaningful probability under the policy's native distribution. Note that we adopt different I in practice for each: the semantic reference uses the rewriting instruction prompt with the privileged trace as shown in Fig. 4, whereas the distributional reference uses a standard question prompt without any rewriting instruction, ensuring the output distribution reflects the model's native generation behavior. Overall, Echo-GRPO replaces $\tilde{y}$ with y_{rewrite} in a way that adheres to the on-policy assumption of GRPO, stabilizing the importance sampling ratios and ensuring that semantically critical tokens remain within the trust region, thereby enabling the model to learn meaningful reasoning components.

4 Experiments

4.1 Experimental Settings

We evaluate three multimodal LLM backbones: InternVL3.5-4B [9], Qwen3-VL-4B [8], and Qwen3-VL-8B, with Qwen3-VL-4B used for all ablations unless noted. We uniformly sample 8 frames per video and generate 6 rollouts per sample, replacing one with the privileged trace for distillation frameworks. All models are optimized using the same training protocol unless otherwise specified. Our training set consists of 2.4K samples from OneThinker-SFT-340K [27], annotated by Seed1.5-VL [10] as teacher policy π_T; privileged traces are used only during training and not at inference time. We use k=5 for DRD and combine accuracy and length rewards. We evaluate on five benchmarks: VideoMMMU [37], MMVU (multiple choice) [38], Video Holmes [39], VSI-Bench [40], and VideoMME [41]. Full hyperparameter details are in the supplement.

4.2 Main results

Tab. 2 reports the main results of VideoEcho-R1 across three backbones (InternVL3.5-4B and Qwen3-VL-4B and Qwen3-VL-8B), comparing against other general training pipelines on five benchmarks spanning general video understanding and reasoning.

Table 2: **Comparison of baselines with general training pipelines across various benchmarks.** We report performance on both Video Reasoning and Video General benchmarks using 8 frames.

Models	VideoMMMU	MMVU(mc)	Video Holmes	VSI-Bench	VideoMME	Avg
InternVL3.5-4B						
Base	43.4	57.4	33.9	37.5	55.4	45.5
+ SFT	54.1	59.0	40.9	42.5	57.4	50.8
+ GRPO	50.4	61.1	40.0	**46.7**	56.5	50.9
+ SFT → GRPO	49.8	61.3	39.4	37.8	57.6	49.2
+ Mixed-Policy GRPO	49.3	58.4	38.8	41.2	55.0	48.5
VideoEcho-R1 (4B)	**55.3**	**65.1**	**41.0**	44.0	**57.9**	**52.7**
Qwen3-VL-4B						
Base	55.2	62.7	33.5	34.7	55.6	48.3
+ SFT	48.2	62.9	40.8	38.2	58.2	49.7
+ GRPO	56.8	68.0	43.3	50.8	**60.7**	55.9
+ SFT → GRPO	57.5	66.2	44.0	42.7	59.6	54.0
+ Mixed-Policy GRPO	42.1	41.9	23.6	35.2	54.5	39.5
VideoEcho-R1 (4B)	**59.1**	**68.8**	**46.7**	**50.9**	**60.7**	**57.2**
Qwen3-VL-8B						
Base	56.3	60.6	36.5	40.2	60.0	50.7
+ SFT	57.0	71.0	42.7	39.8	60.7	54.2
+ GRPO	62.6	71.5	44.7	**51.1**	61.0	**58.2**
+ SFT → GRPO	61.7	69.9	44.6	46.6	60.6	56.7
+ Mixed-Policy GRPO	60.1	71.5	43.1	35.8	57.5	53.6
VideoEcho-R1 (8B)	**63.6**	**72.3**	**45.1**	49.1	**61.1**	**58.2**

Table 3: **Comparison with various reasoning distillation frameworks.** We report performance on both Video Reasoning and Video General benchmarks using 8 frames with Qwen3-VL-4B. $*$ indicates the frameworks originally proposed for LLMs, adopted to our optimization setting with minimal modifications under the same protocols for fair comparison.

Method	VideoMMMU	MMVU(mc)	Video Holmes	VSI-Bench	VideoMME	Avg
RL w/ SFT	52.8	61.0	39.1	45.5	53.5	50.4
RL w/ SFT + Echo	57.2	62.1	40.5	47.1	58.0	53.0
LUFFY*	55.2	62.6	33.3	50.1	58.3	51.9
LUFFY + Echo	56.0	63.8	32.8	50.3	58.3	52.2
OPSD-JSD*	56.1	63.8	36.9	50.8	55.9	52.7
OPSD-IKL*	55.9	63.5	34.1	49.4	56.0	51.8
Echo-GRPO	**59.1**	**68.8**	**46.7**	**50.9**	**60.7**	**57.2**

Table 4: **Comparison of rewriting strategies for Echo-GRPO.** We compare generic paraphrasing, prompt-only student rewriting, and DRD variants using semantic and/or distributional references on Qwen3-VL-4B. Note that generic paraphrasing and prompt-only student rewriting are single-step sentence rewriting of the privilege trace with an external model and the student model, respectively.

Strategy	Sem. Ref.	Dist. Ref.	V-MMMU	MMVU	VHolm	Avg.	Δ
Mixed-Policy	–	–	42.1	41.9	23.6	35.9	–
Generic Paraphrasing	–	–	52.1	64.5	41.4	52.7	+16.8
Prompt-only Student Rewriting	–	–	58.0	67.8	44.9	56.9	+21.0
Semantic-reference only	✓	–	58.5	66.7	45.3	56.8	+20.9
Distributional-reference only	–	✓	57.8	64.0	42.9	54.9	+19.0
DRD (Ours)	✓	✓	**59.1**	**68.8**	**46.7**	**58.2**	**+22.3**

VideoEcho-R1 outperforms on average across backbones and benchmarks. On Qwen3-VL-4B, VideoEcho-R1 achieves an average score of 57.2, surpassing vanilla GRPO by 1.3 points and SFT→GRPO by 3.2 points. On InternVL3.5-4B, VideoEcho-R1 scores 52.7 on average, outperforming GRPO by 1.8 points and SFT→GRPO by 3.5 points. On Qwen3-VL-8B, VideoEcho-R1 reaches 58.2, with consistent gains on reasoning benchmarks such as VideoMMMU (+1.0) and MMVU (+0.8) over GRPO. On VSI-Bench, which rewards exact numerical estimations (*e.g.*, object size, distance), GRPO is competitive or superior. We conjecture that on-policy rollouts produce more consistent numerical reasoning patterns within rollout groups, which is beneficial for tasks requiring

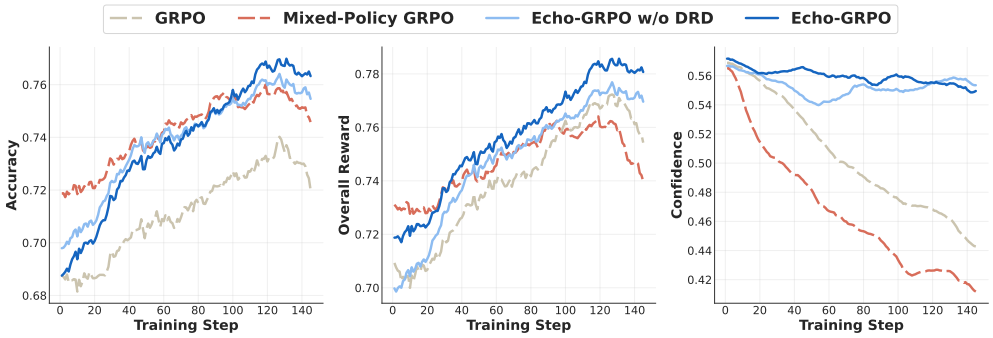


Figure 5: **Training Dynamics of Echo-GRPO.**

fine-grained numerical precision. Notably, this trade-off is partially mitigated when idiolectical rewriting is applied on top of SFT→GRPO with consistent reasoning patterns (see supplement).

Mixed-Policy GRPO empirically degrades reasoning. Across all benchmarks, Mixed-Policy GRPO performs worse than vanilla GRPO with as much as a drop of 16.4 points for Qwen3-VL-4B, a 2.4 points drop for InternVL3.5-4B, and a 4.6 points drop for Qwen3-VL-8B. This empirically confirms our observation in Sec. 3.2, where naive injection of the off-policy privileged reasoning trace hinders effective reasoning distillation.

4.3 Comparison with Reasoning Distillation Frameworks

Tab. 3 compares Echo-GRPO against three representative reasoning distillation frameworks: RL w/ SFT, a GRPO objective with an SFT loss on the privileged trace as ground-truth target, LUFFY, which adopts privileged traces while shaping the gradient for optimization, and the OPSD variants, which distill privileged supervision via fixed-divergence objectives. More details are in the supplement.

Echo-GRPO is the strongest among reasoning distillation methods. Echo-GRPO achieves an average performance of 57.2 across five benchmarks, outperforming RL w/ SFT by 6.8 points, LUFFY by 5.3 points, OPSD-JSD by 4.5 points, and OPSD-IKL by 5.4 points. Among the baselines, OPSD performs competitively, as it adopts divergence objectives, yet still falls short on reasoning-intensive benchmarks. We attribute this gap to its reliance on a single privileged trace per question, which limits the diversity of reasoning patterns. In contrast, Echo-GRPO samples various reasoning traces, exposing the model to a richer set of policy-aligned reasoning paths.

Idiolectal paraphrasing as plug-in module. To assess whether policy-aligned supervision extends beyond Echo-GRPO, we apply our idiolectal paraphrasing to two existing frameworks. RL w/ SFT + Echo improves over RL w/ SFT from 50.4 to 53.0, a gain of 2.6 points. LUFFY + Echo improves over LUFFY from 51.9 to 52.2. The consistent gains across frameworks confirm that policy-aligned supervision with idiolectal paraphrasing is broadly beneficial and is a general plug-in applicable to existing reasoning distillation paradigms, not specific to the GRPO objective.

4.4 Analysis

We analyze various aspects of Echo-GRPO: (1) reduced clipping, (2) training dynamics, (3) training dynamics as a plug-in, (4) generalizability, (5) ablation on the components, (6) ablation on different rewriting strategies, (7) self-paraphrasing beyond RL, and (8) Qualitative results. Unless specified, all analyses use Qwen3-VL-4B.

Reduced clipping on semantically critical tokens. As shown in Fig. 2a (top), unlike Mixed-Policy GRPO, which exhibits a substantial fraction of training samples that contain clipped tokens, suppressing gradient updates on essential reasoning components, Echo-GRPO (blue) alleviates the clip fraction. As visualized in Fig. 2a (bottom), the ratio of semantically important tokens among all clipped tokens is reduced by 4.3 percentage points under Echo-GRPO (67.5%) compared to Mixed-Policy GRPO (71.8%).

Training dynamics of Echo-GRPO. Fig. 5 illustrates the training dynamics of overall accuracy (left), overall reward (middle), and confidence *i.e.*, mean log probability per token (right), In terms of accuracy and reward, Mixed-Policy initially achieves higher accuracy, yet shows a low learning

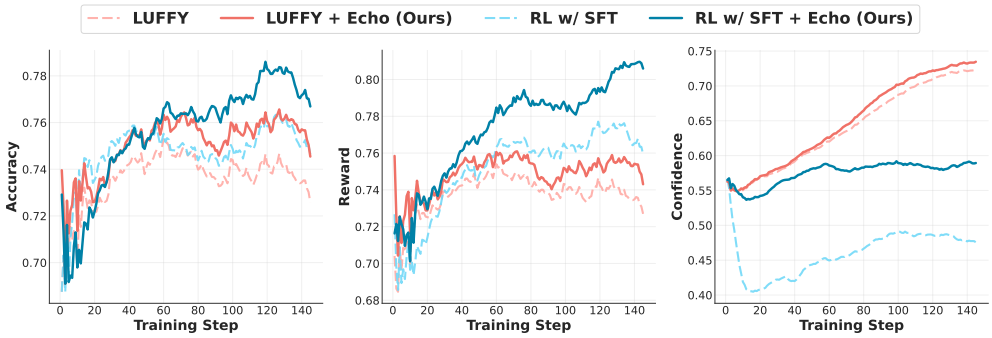


Figure 6: **Training Dynamics of idiolectal paraphrasing as a plug-in.**

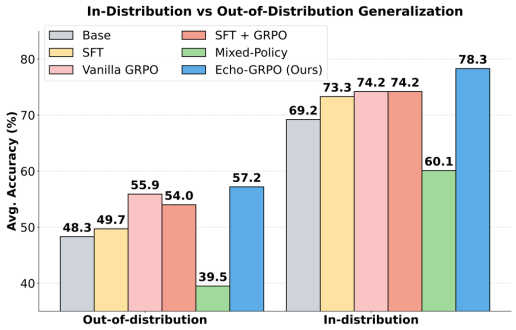


Figure 7: **In-Distribution and Out-of-Distribution Generalization of Echo-GRPO**.

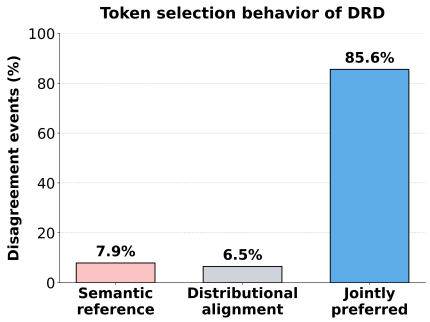


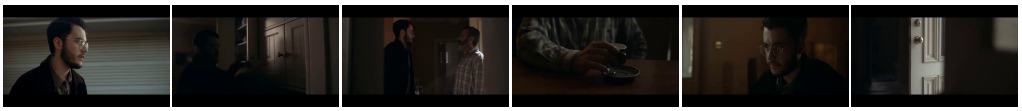
Figure 8: **Token selection behavior of Dual-Reference Decoding**.

slope, while Echo-GRPO demonstrates acceleration with sustained increase, ultimately surpassing all baselines. Also, Echo-GRPO preserves model confidence, while GRPO exhibits a steady decline, and Mixed-Policy collapses in the early stages.

Training dynamics as Plug-in. Fig. 6 reveals that idiolectal paraphrasing is also effective in other reasoning distillation frameworks. LUFFY + Echo and RL w/ SFT + Echo show consistent improvements in confidence, accuracy, and overall reward. Both surpass their respective base frameworks throughout training, indicating idiolectal paraphrasing is broadly applicable and effective.

In- and out-of-distribution generalization. Fig. 7 evaluates generalizability across all methods, where in-distribution refers to the held-out test set aligned with training data and out-of-distribution to general video benchmarks. Echo-GRPO achieves best performance in both regimes (78.3 and 57.2), suggesting that our rewriting enhances reasoning rather than mere mimicking, leading to more robust generalization.

Ablation of rewriting strategy. Tab. 4 compares DRD with simpler rewriting strategies across three representative video reasoning benchmarks. Generic paraphrasing, where an external model (InternVL3.5-4B) rewrites the whole sentence in a single step, already improves Mixed-Policy from 35.9 to 52.7, while prompt-only *student* rewriting that is single-step rewriting by the student model further reaches 56.9, showing that rewriting privileged traces toward the student's distribution is itself beneficial. Beyond this effect, full DRD reaches 58.2, providing an additional 1.3-point gain over prompt-only rewriting. Among the single-reference variants, semantic guidance performs better than distributional guidance (56.8 vs. 54.9), indicating the importance of preserving semantically informative tokens. However, combining both references in DRD consistently improves over semantic-only guidance across all three benchmarks by +0.6, +2.1, and +1.4 points, respectively, yielding the best average performance of 58.2. This further supports jointly incorporating semantic and distributional guidance, also reflected in the training dynamics in Fig. 5. Fig. 8 further shows that, when the two references disagree, DRD selects a token preferred by both references in 85.6% of cases, supporting our product-of-experts formulation.



Question: What is the real relationship between Brandon and the shirt guy?
A. Neighbor **B. Kidnapper and kidnapped person** **C.** deliveryman and customer **D.** colleague
E. Brothers who haven't seen each other for many years **F.** housemate **Ground Truth: B**

Generation with Mixed-Policy GRPO:

So, let's try to figure out the relationship between Brandon and the shirt guy. First, I need to recall what "shirt guy" means. In this context, it's likely a specific person, maybe named Brandon, and the "shirt guy" is another individual. Let's think about the options.
Option A is neighbor. But the question is about their real relationship. Let's check the context. The video is part of a larger series, so we need to look a t the specific details. The answer would depend on the narrative, but since this is a fictional context, we need to infer based on the provided informati on.
Wait, this is part of a very long series, so the answer is likely based on the specific narrative. But without more context, it's hard to say. However, since this is a fictional scenario, the relationship would be determined by the story's rules. So the answer is A (neighbor) for now.
Answer: A

Generation with Echo-GRPO (Ours):

The video shows Brandon and the shirt guy engaging in a series of tense and confrontational interactions. **Brandon is seen entering the house, conf ronting the shirt guy, and later being physically restrained and threatened.** The shirt guy is shown to be in a state of distress and fear, often react ing to Brandon's actions. The interactions suggest a power dynamic where Brandon is in control and the shirt guy is vulnerable. The final scenes show Brandon taking control of the situation, including locking the shirt guy in a room and making him sign a document, which indicates a clear hierarchy a nd dominance. **This relationship is most consistent with a kidnapper and kidnapped person dynamic**, where Brandon is the aggressor and the s hirt guy is the victim.
Answer: B

Figure 9: **Qualitative example.** Mixed-Policy GRPO (top) mimics the structural format of privileged traces, defaulting to narrative speculation ("this is part of a fictional series, so the answer is likely...") and predicting the wrong answer (A: Neighbor). Echo-GRPO (bottom) instead references concrete events from the video (highlighted in green) and reaches the correct answer (B: Kidnapper).

Idiolectal paraphrasing improves SFT beyond RL. Tab. 5 examines whether policy-aligned supervision extends to the SFT setting by replacing the privileged trace $\tilde{y}$ with the rewritten trace y_{rewrite} as the SFT target. For SFT, this improves average performance from 50.6 to 52.6. With SFT→GRPO, performance improves with an average from 55.9 to 58.8. This confirms that idiolectal paraphrasing is broadly effective regardless of the training objective.

Table 5: **Comparison of SFT and SFT→GRPO using offline ($\tilde{y}$) and idiolectal paraphrased traces (y_{rewrite}).**

Method	Priv.	V-MMMU	MMVU	VHolm	Avg.
SFT	$\tilde{y}$	48.2	62.9	**40.8**	50.6
	y_{rewrite}	**56.3**	**64.0**	37.6	**52.6**
SFT→GRPO	$\tilde{y}$	57.5	66.2	44.0	55.9
	y_{rewrite}	**58.2**	**68.8**	**49.5**	**58.8**

Qualitative Results. Fig. 9 compares the generated reasoning trace between the Mixed-Policy GRPO and our Echo-GRPO. Mixed-Policy GRPO fails to ground its reasoning in the video but rather tries to mimic the structural format of the privileged trace, focusing on the narrative speculation *i.e.*, 'this is part of a fictional series, so the answer is likely...'. Echo-GRPO produces a visually grounded reasoning chain, identifying the concrete events (highlighted in green). Hence, the model reaches the correct answer by reasoning with observations rather than imitating the privileged reasoning trace.

5 Conclusion

We identify a failure mode of mixed-policy GRPO, where trust-region clipping suppresses semantically critical tokens. To address this, we propose Echo-GRPO, an idiolectal paraphrasing framework that rewrites privileged reasoning into the model's native distribution via Dual-Reference Decoding. Our VideoEcho-R1 achieves consistent improvements across multiple backbones and benchmarks, and generalizes as a plug-in module for both reinforcement learning and supervised fine-tuning.

Acknowledgments and Disclosure of Funding

This work was supported in part by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grants funded by the Korea government (MSIT) Accurate and Safe Multimodal, Multilingual Personalized AI Tutors (No. RS-2024-00443251, 30%), LG AI STAR Talent Development Program for Leading Large-Scale Generative AI Models in the Physical AI Domain (No. RS-2025-25442149, 10%). This work was also supported in part by the InnoCORE

program of the Ministry of Science and ICT (N10250156, 30%) and the National Supercomputing Center with supercomputing resources including technical support (KSC-2025-CRE-0506, 30%). The authors declare no competing interests.

References

[1] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

[2] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

[3] Kimi Team, Angang Du, Bofei Gao, Bowei Xing, Changjiu Jiang, Cheng Chen, Cheng Li, Chenjun Xiao, C Du, C Liao, et al. Kimi k1. 5: Scaling reinforcement learning with llms, 2025. *URL https://arxiv.org/abs/2501.12599*, 118, 2025.

[4] Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. Openai o1 system card. *arXiv preprint arXiv:2412.16720*, 2024.

[5] Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, et al. Dapo: An open-source llm reinforcement learning system at scale. In *NeurIPS*, 2025.

[6] Jianhao Yan, Yafu Li, Zican Hu, Zhi Wang, Ganqu Cui, Xiaoye Qu, Yu Cheng, and Yue Zhang. Learning to reason under off-policy guidance. In *NeurIPS*, 2025.

[7] Yunheng Li, Jing Cheng, Shaoyong Jia, Hangyi Kuang, Shaohui Jiao, Qibin Hou, and Ming-Ming Cheng. Tempsamp-r1: Effective temporal sampling with reinforcement fine-tuning for video llms. In *NeurIPS*, 2025.

[8] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

[9] Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025.

[10] ByteDance Seed Team. Seed1.5-vl technical report. *arXiv preprint arXiv:2505.07062*, 2025.

[11] Dohwan Ko, Ji Soo Lee, Woo-Young Kang, Byungseok Roh, and Hyunwoo J Kim. Large language models are temporal and causal reasoners for video question answering. In *EMNLP*, 2023.

[12] Dohwan Ko, Ji Soo Lee, Minhyuk Choi, Zihang Meng, and Hyunwoo J Kim. Bidirectional likelihood estimation with multi-modal large language models for text-video retrieval. In *ICCV*, 2025.

[13] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. In *NeurIPS*, 2022.

[14] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *NeurIPS*, 2023.

[15] Ke Zhu, Liang Zhao, Zheng Ge, and Xiangyu Zhang. Self-supervised visual preference alignment. In *ACMMM*, 2024.

[16] Zhaolin Gao, Jonathan D. Chang, Wenhao Zhan, Owen Oertell, Gokul Swamy, Kianté Brantley, Thorsten Joachims, J. Andrew Bagnell, Jason D. Lee, and Wen Sun. Rebel: Reinforcement learning via regressing relative rewards. In *NeurIPS*, 2024.

[17] Yiyang Zhou, Chenhang Cui, Rafael Rafailov, Chelsea Finn, and Huaxiu Yao. Aligning modalities in vision large language models via preference fine-tuning. *arXiv preprint arXiv:2402.11411*, 2024.

[18] Byungoh Ko, Jinyoung Park, Jongha Kim, Jeehye Na, Jaewon Cho, and Hyunwoo J Kim. Context blindness in dpo: Mitigating object hallucination in mllms via context-calibrated preference optimization. In *ECCV*, 2026.

[19] Ji Soo Lee, Byungoh Ko, Jaewon Cho, Howoong Lee, Jaewoon Byun, and Hyunwoo J Kim. Captioning for text-video retrieval via dual-group direct preference optimization. In *EMNLP Findings*, 2025.

[20] Ji Soo Lee, Jongha Kim, Jeehye Na, Jinyoung Park, and Hyunwoo J Kim. Vidchain: Chain-of-tasks with metric-based direct preference optimization for dense video captioning. In *AAAI*, 2025.

[21] Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. Understanding r1-zero-like training: A critical perspective. In *COLM*, 2025.

[22] Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, Rui Men, An Yang, et al. Group sequence policy optimization. *arXiv preprint arXiv:2507.18071*, 2025.

[23] Jinyoung Park, Jeehye Na, Jinyoung Kim, and Hyunwoo J Kim. Deepvideo-r1: Video reinforcement fine-tuning via difficulty-aware regressive grpo. In *NeurIPS*, 2025.

[24] Youssef Mroueh, Nicolas Dupuis, Brian Belgodere, Apoorva Nitsure, Mattia Rigotti, Kristjan Greenewald, Jiri Navratil, Jerret Ross, and Jesus Rios. Revisiting group relative policy optimization: Insights into on-policy and off-policy training. *arXiv:2505.22257*, 2025.

[25] Siyan Zhao, Zhihui Xie, Mengchen Liu, Jing Huang, Guan Pang, Feiyu Chen, and Aditya Grover. Self-distilled reasoner: On-policy self-distillation for large language models. *arXiv preprint arXiv:2601.18734*, 2026.

[26] Wenxuan Huang, Bohan Jia, Zijie Zhai, Shaosheng Cao, Zheyu Ye, Fei Zhao, Zhe Xu, Xu Tang, Yao Hu, and Shaohui Lin. Vision-r1: Incentivizing reasoning capability in multimodal large language models. In *ICLR*, 2026.

[27] Kaituo Feng, Manyuan Zhang, Hongyu Li, Kaixuan Fan, Shuang Chen, Yilei Jiang, Dian Zheng, Peiwen Sun, Yiyuan Zhang, Haoze Sun, et al. Onethinker: All-in-one reasoning model for image and video. In *CVPR*, 2026.

[28] Xinhao Li, Ziang Yan, Desen Meng, Lu Dong, Xiangyu Zeng, Yinan He, Yali Wang, Yu Qiao, Yi Wang, and Limin Wang. Videochat-r1: Enhancing spatio-temporal perception via reinforcement fine-tuning. *arXiv preprint arXiv:2504.06958*, 2025.

[29] Ye Wang, Boshen Xu, Zihao Yue, Zihan Xiao, Ziheng Wang, Liang Zhang, Dingyi Yang, Wenxuan Wang, and Qin Jin. Timezero: Temporal video grounding with reasoning-guided lvlm. *arXiv preprint arXiv:2503.13377*, 2025.

[30] Yi Chen, Yuying Ge, Rui Wang, Yixiao Ge, Lu Qiu, Ying Shan, and Xihui Liu. Exploring the effect of reinforcement learning on video understanding: Insights from seed-bench-r1. *arXiv preprint arXiv:2503.24376*, 2025.

[31] Ye Wang, Ziheng Wang, Boshen Xu, Yang Du, Kejun Lin, Zihan Xiao, Zihao Yue, Jianzhong Ju, Liang Zhang, Dingyi Yang, et al. Time-r1: Post-training large vision language model for temporal video grounding. *arXiv preprint arXiv:2503.13377*, 2025.

[32] Xingjian Zhang, Siwei Wen, Wenjun Wu, and Lei Huang. Tinyllava-video-r1: Towards smaller lmms for video reasoning. *arXiv preprint arXiv:2504.09641*, 2025.

[33] Shuming Liu, Mingchen Zhuge, Changsheng Zhao, Jun Chen, Lemeng Wu, Zechun Liu, Chenchen Zhu, Zhipeng Cai, Chong Zhou, Haozhe Liu, et al. Videoauto-r1: Video auto reasoning via thinking once, answering twice. In *CVPR*, 2026.

[34] Seohyun Lee, Seoung Choi, Dohwan Ko, Jongha Kim, and Hyunwoo J. Kim. Videosearch-r1: Iterative video retrieval and reasoning via soft query refinement. In *ECCV*, 2026.

[35] Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, Junfei Wu, Xiaoying Zhang, Benyou Wang, and Xiangyu Yue. Video-r1: Reinforcing video reasoning in mllms. In *NeurIPS*, 2025.

[36] Huajie Tan, Yuheng Ji, Xiaoshuai Hao, Xiansheng Chen, Pengwei Wang, Zhongyuan Wang, and Shanghang Zhang. Reason-rft: Reinforcement fine-tuning for visual reasoning of vision language models. In *NeurIPS*, 2025.

[37] Kairui Hu, Penghao Wu, Fanyi Pu, Wang Xiao, Yuanhan Zhang, Xiang Yue, Bo Li, and Ziwei Liu. Video-mmmu: Evaluating knowledge acquisition from multi-discipline professional videos. *arXiv preprint arXiv:2501.13826*, 2025.

[38] Yilun Zhao, Haowei Zhang, Lujing Xie, Tongyan Hu, Guo Gan, Yitao Long, Zhiyuan Hu, Weiyuan Chen, Chuhan Li, Zhijian Xu, et al. Mmvu: Measuring expert-level multi-discipline video understanding. In *CVPR*, 2025.

[39] Junhao Cheng, Yuying Ge, Teng Wang, Yixiao Ge, Jing Liao, and Ying Shan. Video-holmes: Can mllm think like holmes for complex video reasoning? *arXiv preprint arXiv:2505.21374*, 2025.

[40] Jihan Yang, Shusheng Yang, Anjali W Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in space: How multimodal large language models see, remember, and recall spaces. In *CVPR*, 2025.

[41] Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *CVPR*, 2025.

[42] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, 2019.

[43] Yaowei Zheng, Junting Lu, Shenzhi Wang, Zhangchi Feng, Dongdong Kuang, and Yuwen Xiong. Easyr1: An efficient, scalable, multi-modality rl training framework. `https://github.com/hiyouga/EasyR1`, 2025.

[44] Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, Zhangchi Feng, and Yongqiang Ma. Llamafactory: Unified efficient fine-tuning of 100+ language models. In *ACL*, 2024.

[45] Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *ACM SIGOPS 29th Symposium on Operating Systems Principles*, 2023.

[46] Mathematical Association of America. American invitational mathematics examination (AIME) 2024, 2024. AIME I and AIME II competition problems.

[47] Mathematical Association of America. American invitational mathematics examination (AIME) 2025, 2025. AIME I and AIME II competition problems.

[48] Jasper Dekoninck, Nikola Jovanović, Tim Gehrunger, Kári Rögnvaldsson, Ivo Petrov, Chenhao Sun, and Martin Vechev. Beyond benchmarks: Matharena as an evaluation platform for mathematics with llms, 2026.

[49] Etash Guha, Ryan Marten, Sedrick Keh, Negin Raoof, Georgios Smyrnis, Hritik Bansal, Marianna Nezhurina, Jean Mercat, Trung Vu, Zayne Sprague, Ashima Suvarna, Benjamin Feuer, Liangyu Chen, Zaid Khan, Eric Frankel, Sachin Grover, Caroline Choi, Niklas Muennighoff, Shiye Su, Wanjia Zhao, John Yang, Shreyas Pimpalgaonkar, Kartik Sharma, Charlie Cheng-Jie Ji, Yichuan Deng, Sarah Pratt, Vivek Ramanujan, Jon Saad-Falcon, Jeffrey Li, Achal Dave, Alon Albalak, Kushal Arora, Blake Wulfe, Chinmay Hegde, Greg Durrett, Sewoong Oh, Mohit Bansal, Saadia Gabriel, Aditya Grover, Kai-Wei Chang, Vaishaal Shankar, Aaron Gokaslan, Mike A. Merrill, Tatsunori Hashimoto, Yejin Choi, Jenia Jitsev, Reinhard Heckel, Maheswaran Sathiamoorthy, Alexandros G. Dimakis, and Ludwig Schmidt. Openthoughts: Data recipes for reasoning models, 2025.

[50] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

A Experimental Details

We adopt three multimodal large language model backbones: InternVL3.5-4B [9], Qwen3-VL-4B [8], and Qwen3-VL-8B [8]. All models are trained with uniformly sampled 8 video frames per input. We set the learning rate to 9×10^{-7} and implement our framework in PyTorch [42], building on the EasyR1 [43] library for post-training and the LLaMAFactory [44] framework for supervised fine-tuning. For rollout generation and inference, we utilize vLLM [45]. For optimization, we adopt a per-step rollout batch size of 16. Within each update, we split the rollout batch into minibatches of size 4, which defines the actor-update minibatch size. Training is conducted on NVIDIA RTX PRO 6000 Blackwell Max-Q GPUs.

We sample 2.4k training instances from the OneThinker-SFT-340K [27] dataset, focusing on video-centric and multiple-choice examples annotated by Seed1.5-VL [10]. Following our training protocol, we generate $G = 6$ rollouts per sample, where one rollout is replaced with the distilled trace when distillation is applied. We omit the KL divergence term during optimization. For Dual-Reference Decoding, we use a top-k value of $k = 5$. All evaluations are conducted with a decoding temperature of 0.0 to ensure deterministic outputs. In addition, we adopt LLM-based tools for the generation of the prompt and for correcting grammatical errors in the writing.

The idiolectal paraphrasing prompt is in Fig. 4 of main. The instruction `Do not mention the reference` is included to prevent the model from explicitly citing or repeating the privileged trace in its response. Without this constraint, we observe that the model tends to produce outputs that directly reference the provided trace, such as "according to the reference" or "as mentioned above", effectively copying its structure rather than genuinely paraphrasing its semantics into the policy's native distribution. While instruction prevents surface-level copying of the privileged trace, it does not guarantee that generated tokens lie within the student policy's native distribution with requires our Dual-Reference Decoding.

B Baselines

In this section, we describe the baseline methods used for comparison.

SFT trains the model to imitate target reasoning traces using a standard cross-entropy objective. We adopt LoRA for parameter-efficient fine-tuning, which we found to yield strong performance compared to full-parameter training in our setting.

GRPO [2] is an on-policy reinforcement learning algorithm that optimizes the model using group-wise normalized advantages computed over multiple sampled rollouts per prompt. The policy is updated using reward-weighted likelihood ratios without relying on explicit supervision.

SFT $\rightarrow$ GRPO adopts a two-stage training strategy. The model is first initialized via SFT (cold start) and subsequently optimized using GRPO. For fair comparison we use 0.8K samples for SFT initialization and the remaining data for GRPO training.

Mixed-Policy GRPO [6] incorporates off-policy supervision by injecting ground-truth reasoning traces during RL training. Specifically, one of the sampled rollouts is replaced with a ground-truth trace, while the remaining rollouts are generated from the current policy. This introduces a controlled off-policy signal within the GRPO framework.

RL w/ SFT jointly optimizes supervised and reinforcement learning objectives, which was introduced in [6]. Privileged reasoning is trained with an SFT loss, while the remaining sampled rollouts are optimized using the RL objective. This hybrid objective enables simultaneous learning from explicit supervision and reward-driven exploration.

LUFFY [6] is an off-policy reinforcement learning framework that leverages external reasoning traces to guide policy optimization. It combines off-policy demonstrations with on-policy rollouts and applies regularized importance weighting to stabilize training under distribution mismatch, improving generalization beyond purely on-policy methods.

OPSD [25] is a training paradigm in which the model learns from its own generated trajectories, adopting a single model as teacher (with privileged trace) and student by treating them as supervision signals. It keeps the privileged trace as a distributional target in the loss function, training instead over student-generated rollouts via per-token KL divergence, while adopting only one trace for rollout.

C Evaluation Benchmarks

We evaluate our method on a diverse suite of video question answering benchmarks that cover both general video understanding and reasoning-intensive QA scenarios.

Video-MMMU [37] evaluates the capacity of MLLMs to acquire knowledge and perform sophisticated reasoning. Formulated around the human cognitive progression of perceiving, comprehending, and adapting, this dataset features 900 questions paired with 300 highly specialized videos. The queries are manually crafted and distributed across six distinct fields—Medicine, Engineering, Art, Business, Science, and Humanities—requiring models to synthesize temporal visual cues with deep domain expertise.

MMVU (MC) [38] targets expert-grade, knowledge-dense video comprehension. It encompasses 1,529 videos sourced from professional domains, accompanied by 3,000 expertly curated QA pairs. The benchmark is broadly divided into four primary branches (Healthcare, Engineering, Science, and Humanities & Social Sciences), which are further subdivided into 27 specific subjects. In alignment with prior methodologies, our evaluation utilizes the multiple-choice variant, tasking models with synthesizing contextual evidence to deduce the correct candidate.

Video-Holmes [39] is a reasoning-intensive benchmark designed to assess the complex, multi-step deductive capabilities of MLLMs. It comprises 1,837 questions sourced from 270 expertly annotated suspense short films, with durations ranging from 1 to 5 minutes. Diverging from traditional datasets that provide explicit context, Video-Holmes is structured around an "active seeking" paradigm. It evaluates models across seven challenging tasks—including temporal causal inference, intention and motive chaining, and physical anomaly reasoning—requiring the model to actively locate, temporally connect, and synthesize scattered visual evidence to resolve complex causal relationships.

VSIBench [40] is designed to gauge visual-spatial intelligence and numerical inference. Utilizing a collection of 288 authentic videos and over 5,000 QA pairs, it challenges models across three distinct reasoning paradigms: spatiotemporal tracking, configuration analysis, and measurement estimation. This involves granular objectives such as estimating absolute or relative distances, counting objects, and determining appearance order, making it an ideal testbed for rigorous structural reasoning.

Video-MME [41] offers a holistic assessment of overarching video comprehension. It contains 2,700 QA pairs tied to 900 videos that exhibit significant diversity in both content—spanning 30 sub-categories within 6 major domains—and duration. To rigorously test temporal scalability, the benchmark categorizes videos into short (under 2 minutes), medium (4 to 15 minutes), and long (30 to 60 minutes) intervals. To ensure a true measure of raw visual-semantic processing, we report the mean accuracy across all duration brackets without providing textual subtitles.

Overall, these benchmarks jointly assess the model's ability to perform reasoning-driven video question answering, ranging from general understanding to structured, multi-step inference and numerical reasoning.

D Additional Analysis

This section provides supplementary analyses to complement the main experimental results. We report full benchmark results for the SFT→GRPO pipeline with idiolectal paraphrased traces, and analyze the sensitivity of Dual-Reference Decoding to the top-k hyperparameter. Together, these analyses further validate the generality and robustness of Echo-GRPO across training objectives and decoding configurations.

D.1 Full results of SFT→GRPO with Idiolectal paraphrased rewriting.

Table 6: **Results of SFT→GRPO using offline ($\tilde{y}$) and idiolectal paraphrased traces (y_{rewrite}).**

Method	Priv. trace	V-MMMU	MMVU	VHolmes	VSI-Bench	VideoMME	Avg.
SFT→GRPO	$\tilde{y}$	57.5	66.2	44.0	42.7	59.6	54.0
	y_{rewrite}	**58.2**	**68.8**	**49.5**	**51.7**	**60.6**	**57.8**

Tab. 6 reports the full benchmark results of SFT→ GRPO for Qwen3-VL-4B using offline privileged traces ($\tilde{y}$) versus idiolectal paraphrased traces (y_{rewrite}). Replacing $\tilde{y}$ with y_{rewrite} yields consistent improvements across all five benchmarks, with an average gain of 3.8 points (54.0 → 57.8). Notably, the gains are particularly pronounced on reasoning-intensive benchmarks such as Video Holmes (+5.5) and VideoMMMU (+0.7). Of particular interest is VSI-Bench, where idiolectal paraphrasing achieves a substantial improvement of 9.0 points (42.7 → 51.7), comparable to Echo-GRPO (50.9), suggesting that the SFT warm-start provides a stable numerical reasoning foundation that mitigates the precision trade-off observed in vanilla Echo-GRPO. These results confirm that the benefit of policy-aligned supervision via idiolectal paraphrasing is not specific to the GRPO objective, but extends to the SFT → GRPO pipeline as well.

D.2 Sensitivity to top-k in Dual-Reference Decoding.

Table 7: **Sensitivity of top-k in DRD.**

top-k	VideoMMMU	MMVU	VideoHolmes	Avg.
1	58.1	67.9	45.1	57.0
5	**59.1**	**68.8**	**46.7**	**58.2**
10	57.1	65.0	38.2	53.4

A key hyperparameter of DRD is the size k of the semantic reference candidate set $V_k{}^{(t)}$. Tab. 7 reports performance across $k \in \{1, 5, 10\}$ on VideoMMMU, MMVU, and Video Holmes. At $k = 1$, DRD enforces strict semantic fidelity by restricting token selection to the single most probable token under the semantic reference, yet achieves a competitive average of 57.0, confirming that semantic grounding alone is beneficial. At $k = 5$, DRD achieves the best average performance of 58.2, striking the optimal balance between semantic faithfulness and distributional alignment. At $k = 10$, performance drops substantially to 53.4, suggesting that relaxing the semantic constraint too much allows the distributional reference excessive freedom, leading to semantic drift from the privileged trace and ultimately hurting reasoning quality. These results confirm that $k = 5$ is the sweet spot, where the candidate set is large enough to allow distributionally natural token choices while remaining constrained enough to preserve the semantic content of the privileged trace.

D.3 Domain generalization of Echo-GRPO

Table 8: **Effect of Echo-GRPO on text reasoning benchmarks.**

Method	**AIME24**	**AIME25**	**HMMT25**	**Avg.**	**Δ**
Base	16.7	26.7	10.0	17.8	–
Mixed-Policy	36.7	40.0	16.7	31.1	+13.3
GRPO	40.0	40.0	20.0	33.3	+15.5
Echo-GRPO	**53.3**	**46.7**	**23.3**	**41.1**	**+23.3**

To evaluate whether Echo-GRPO generalizes beyond video question answering, in Tab 8, we conduct experiments on three widely-used for text reasoning benchmarks: AIME24 [46], AIME25 [47], and HMMT25 [48]. Following OPSD [25], we train on a 5K subsample of the OpenThought [49] dataset. We train and evaluate Qwen3-4B [50] in non-thinking mode, with a maximum sequence length of 32,768 and temperature 0.0. As shown below, Echo-GRPO achieves the best performance across all three benchmarks, improving over the base model by +23.3 points on average, over Mixed-Policy by +10.0, and over GRPO by +7.8. These results indicate that Echo-GRPO's policy-aligned rewriting extends beyond the video domain to text-based math reasoning, supporting the generality of the approach.

D.4 Scale-up Experiment

We also conducted a scaled-up experiment with approximately four times as large training set that is a total of 9k samples. As presented in Tab 9, the Mixed-policy optimization shows improvement

Table 9: **Performance and clipping statistics during scaled-up training.** The average performance covers three major benchmarks: VideoMMMU, MMVU, and Video-Holmes. Average Clip presents the average percentage of samples clipped throughout the steps.

Method	**Avg. Perf. (↑)**	**Avg. Clip (↓) (%)**	**0–100**	**100–200**	**200–300**	**300–400**	**400–500**	**500+**
Base	50.5	–	–	–	–	–	–	–
Mixed-Policy	51.7 (+1.2)	26.1	11.2	9.0	21.5	44.6	42.1	28.2
GRPO	56.4 (+5.9)	7.8	9.6	8.4	8.2	6.9	7.7	5.7
Echo-GRPO	**58.2** (+7.7)	**7.9**	**10.2**	**8.6**	**7.3**	**7.5**	**8.1**	**5.5**

compared to the original of which it yields an average of 51.7 compared to 35.9 on the three major benchmarks of VideoMMMU, MMVU, and Video-Holmes. However, specifically for Mixed-policy we observe that distribution mismatch still accumulates during training: the clipping ratio rises sharply around step 200 and reaches an average of 26.1% over the later training stages. In contrast, Echo-GRPO remains stable as the policy evolves, maintaining an average clipping ratio of only 7.9%. It also outperforms Mixed-Policy by +6.5 points on average and vanilla GRPO by +1.8 points. These results rule out the possibility that our previous findings were driven by the small training set. Even at a larger scale, Echo-GRPO provides more stable video-reasoning optimization by reducing the clipping of important tokens while keeping the rewritten trajectories aligned with the evolving student distribution.

E Computation Analysis of DRD

Table 10: **Performance vs. offline construction cost.** Avg. denotes the mean performance across VideoMMMU, MMVU, and Video-Holmes. Idiolectal paraphrasing is performed once during data construction; it introduces no additional training or inference-time overhead.

Strategy	Avg.	Δ vs. GRPO	Offline cost†
GRPO	56.0	–	–
Echo-GRPO w/o DRD	56.9	+0.9	9.4 ms/token
Echo-GRPO w/ DRD	**58.2**	**+2.2**	163.9 ms/token

†One-time cost for constructing and caching the paraphrased training traces. Training and deployment incur no additional inference overhead.

Tab. 10 reports the average performance across three video reasoning benchmarks and the offline construction cost of each idiolectal paraphrasing strategy. Echo-GRPO without DRD improves over vanilla GRPO by 0.9 points with lightweight single-pass paraphrasing, while incorporating DRD yields a further 1.3-point improvement, for a total gain of 2.2 points over GRPO. DRD requires additional computation during paraphrase generation because both semantic and distributional references are evaluated during decoding. This computation is restricted to offline data construction: the resulting traces are generated once and cached before training. Consequently, DRD introduces no additional decoding overhead during subsequent training or deployment. We therefore view the construction overhead as a one-time preprocessing cost that is amortized across training, analogous to offline dataset curation in supervised fine-tuning pipelines.

F Broader Impact and Limitations

F.1 Broader Impact

We propose Echo-GRPO, an idiolectal paraphrasing framework for reasoning distillation in videoLLMs, instantiated as VideoEcho-R1. We believe Echo-GRPO itself does not introduce new negative societal impacts. However, as VideoEcho-R1 is built upon pretrained multimodal language models, it may inherit biases present in pretraining data, potentially generating outputs that reflect stereotypes related to race, religion, culture, or gender. Careful deployment aligned with responsible AI principles is necessary.

F.2 Limitations

Echo-GRPO introduces mild computational overhead during preprocessing due to Dual-Reference Decoding's two forward passes per step, though rewritten traces are cached prior to training and incur no inference-time cost. Additionally, the quality of idiolectal paraphrased traces is bounded by the teacher policy's reasoning quality, and performance gains may be reduced on tasks requiring fine-grained numerical precision, as discussed in Sec. 4.2. Finally, as VideoEcho-R1 is fine-tuned on top of large pretrained models, potential overlap between pretraining content and evaluation benchmarks introduces a risk of implicit data leakage.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Presented in Abstract and Introduction.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: In the appendix as separate section.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

Justification: This paper does not include theoretical results. The contributions are algorithmic and empirical, validated through experiments across multiple backbones and benchmarks.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Present in the Experiments Section and the Supplement.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in the supplemental material?

Answer: [Yes]

Justification: We present implementation details in the supplement and the code.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Present in the Experiments and the Supplement.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Report results of one-time experiment with temperature 0.0.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)

- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Explained in the supplement.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Yes. Reviewed all.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Explained in the supplement as a separate section.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [Yes]

    Justification: In the supplement.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: In the supplement.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: Details in the Experiments and the Supplement.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: No crowdsourcing experiment required.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: Not applicable.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: In the supplement.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.