# Supplementary for Video Reasoning Without Training

Deepak Sridhar[1,2*] Kartikeya Bhardwaj[1*] Jeya Pradha Jeyaraj[1] Nuno Vasconcelos[2]
Ankita Nayak[1] Harris Teague[1]
[1]Qualcomm AI Research[†] [2]University of California, San Diego
desridha@ucsd.edu, kbhardwa@qti.qualcomm.com

## A. Theoretical Analysis: Bounding Entropy under Switching Loss

Let the vocabulary size be $n = |\mathcal{V}|$. At generation step $t$, the model (with value-cache controller parameters $\Delta V$) produces logits $z_t \in \mathbb{R}^n$ and probabilities

$$p_t(\Delta V) = \mathrm{softmax}(z_t(\Delta V)), \qquad \sum_i p_t^i = 1.$$

The Shannon entropy of this distribution is

$$H_t(\Delta V) := -\sum_{i=1}^{n} p_t^i(\Delta V) \log p_t^i(\Delta V),$$

and its exponential moving average (EMA) is

$$H_t^{ema} = \beta H_{t-1}^{ema} + (1-\beta) H_t, \quad \beta \in (0,1).$$

The Entropy Switching Loss at optimization step $k$ is

$$\mathcal{L}_{switch}(\Delta V) = -\alpha_k H_k(\Delta V),$$

where the coefficient $\alpha_k \in \{-1, +1\}$ is defined as

$$\alpha_k = \begin{cases} +1 & \text{if } H_k^{ema} \geq H_{peak}^{ema}, \\ -1 & \text{otherwise,} \end{cases}$$

with $H_{peak}^{ema}$ denoting the maximum EMA value observed before step $k$.

**Assumptions.** We make the following assumptions:

1. Logits $z_t(\Delta V)$ are smooth in $\Delta V$, and $\partial z_t / \partial \Delta V$ is bounded. From equation 1, $\mathbb{V}_L^{new} = \frac{\mathbb{V}_L + \Delta V}{||\mathbb{V}_L + \Delta V||} \cdot ||\mathbb{V}_L||$. So, $\partial z_t / \partial \Delta V$ being bounded is a valid assumption because the update to value cache is bounded by the normalization factor which only provides a directional update.
2. The optimizer uses a bounded step size (learning rate) $\eta > 0$ and updates are sufficiently small per step (i.e., standard stochastic gradient/Lipschitz assumptions).
3. Vocabulary size is finite, hence $H_t \in [0, \log n]$ for all $t$.

**Preliminaries.** Differentiating the entropy with respect to logits yields

$$\nabla_z H = -J_p^\top (\mathbf{1} + \log p),$$

where $J_p = \partial p / \partial z$ is the softmax Jacobian. Since $\|J_p\|$ is bounded and $\mathbf{1} + \log p$ is finite (as $p_i \in (0,1]$), we obtain

$$\|\nabla_{\Delta V} H\| \leq C$$

for some constant $C$.

**Proposition 1** (Bounded entropy updates)**.** *Under the assumptions above, one gradient step of size $\eta$ on $\mathcal{L}_{switch}$ changes entropy by at most*

$$|H_{t+1} - H_t| \leq \eta C + o(\eta),$$

*and the process $\{H_t\}$ remains in the compact interval $[0, \log n]$. Here, $o(\eta)$ denotes the higher-order terms from the Taylor expansion of $H(\Delta V)$ around the current iterate.*

*Proof.* First, the gradient of entropy with respect to controller parameters is

$$\nabla_{\Delta V} H_k(\Delta V) = \frac{\partial H_k}{\partial z_k} \frac{\partial z_k}{\partial \Delta V}.$$

**Bounding $\nabla_z H_k$.** For softmax probabilities bounded away from 0 and 1, the Jacobian $J_p = \partial p_k / \partial z_k$ satisfies $\|J_p\|_2 \leq 1$. Moreover, the entropy gradient w.r.t. logits is

$$\nabla_z H_k = -J_p^\top (\mathbf{1} + \log p_k),$$

and $\|\mathbf{1} + \log p_k\|_2 \leq \sqrt{n} \max_i |1 + \log p_k^i| \leq C_1$ for some constant $C_1$ depending on $n$ and $\epsilon$ (the lower bound on softmax probabilities). Therefore,

$$\|\nabla_z H_k\|_2 \leq C_1.$$

**Bounding $\nabla_{\Delta V} H_k$.** Since $z_k$ is $L_z$-Lipschitz in $\Delta V$,

$$\|\nabla_{\Delta V} H_k\|_2 = \|\nabla_z H_k \cdot \partial z_k / \partial \Delta V\|_2 \leq C_1 L_z := L_H.$$

[*]Equal contribution. Work done when Deepak Sridhar was an intern at Qualcomm AI Research.

[†]Qualcomm AI Research is an initiative of Qualcomm Technologies, Inc.

**Bounding one gradient step.** A single gradient step updates the controller:

$$\Delta V \leftarrow \Delta V + \eta \alpha_k \nabla_{\Delta V} H_k.$$

Using the Lipschitz property of $H_k$ w.r.t $\Delta V$,

$$|H_k(\Delta V+\eta\alpha_k \nabla_{\Delta V} H_k) - H_k(\Delta V)| \leq \eta \|\nabla_{\Delta V} H_k\|_2 \leq \eta L_H.$$

**Global bounds.** Since $H_k \in [0, \log n]$ by definition, this step-size bound guarantees the entropy remains in $[0, \log n]$ after each update.

□

**Proposition 2** (EMA smoothing bounds oscillations)**.** *For $\beta \in (0,1)$ close to 1, the EMA acts as a low-pass filter:* (*i*) *it attenuates high-frequency fluctuations of $H_t$,* (*ii*) *delays the attainment of entropy maxima, and* (*iii*) *enforces bounded oscillations by switching $\alpha_k$ to $-1$ once a new global EMA maximum is reached.*

*Proof.* (*i*) The recursion $H_t^{ema} = \beta H_{t-1}^{ema} + (1-\beta) H_t$ is a causal low-pass filter, suppressing fast oscillations. (*ii*) Because $H^{ema}$ averages over past values, peaks in $H_t$ appear later and at lower amplitude in $H^{ema}$, creating delayed switching. (*iii*) Once $H^{ema}$ reaches a global maximum, $\alpha = -1$, turning the loss into an entropy-minimization objective. This guarantees the entropy trajectory descends after each peak, bounding the amplitude of oscillations. □

**Discussion.** The trivial upper bound $H_t \leq \log n$ already prevents unbounded entropy; Proposition 1 strengthens this by showing the optimization dynamics cannot instantaneously jump arbitrarily close to $\log n$ provided the learning rate is small and gradients are bounded. In practice, this prevents pathological "entropy blow-ups" during optimization. EMA smoothing makes the switching decision depend on sustained increases in entropy rather than on single noisy spikes. These results imply that the Entropy Switching Loss enforces *bounded micro-cycles* of exploration and exploitation: entropy increases are promoted only when sustained (captured by $H^{ema}$), while decreases are enforced once a peak is reached. This yields lower and delayed entropy maxima, consistent with the empirical patterns of stronger reasoning models.

Concurrent work, Top-H [1], formalizes entropy bounds in the decoding step by solving (approximately) an entropy-constrained minimization problem that upper-bounds the randomness of the truncated distribution while keeping divergence from the model distribution small. Our approach uses a complementary perspective: rather than imposing a hard constraint on the sampling distribution at each decoding step, we *optimize the controller* so that the model's intrinsic token distributions themselves enter phases of controlled exploration and exploitation (via maximizing/minimizing $H$ at different times). The EMA-based switching mirrors the time-adaptive, entropy-aware thresholds used in Top-H while operating *inside* the model (controller optimization) rather than as an external truncation rule. Empirically and theoretically, both approaches rely on the same fundamental fact: *entropy is a natural, bounded quantity* that can be used as a control signal to trade-off diversity and consistency in generation.

# B. Implementation Details

**Hyperparameters.** AdamW optimizer is used to update the controller with no weight decay. A step size of $k = 4$ is used as default unless otherwise specified and the best accuracy is reported over a grid search of 10 learning rates from `5e-5` to `5e-4`. The gradient norm of the value-cache controller is clipped to 1.0. We used $\beta = 0.98$ smoothing factor for EMA.

**Evaluation.** Classification accuracy is computed as the proportion of correct answers to the multiple-choice QA. Mean Relative Accuracy measures the proportion of predictions whose relative error falls below a series of thresholds ranging from 0.5 to 0.95. The final score is the average accuracy across all thresholds. For VSI-Bench, we report both classification accuracy and MRA individually, as well as their average. To compute the overall average accuracy across all six datasets, we divide by seven, treating the two scores from VSI-Bench separately in addition to the other datasets. When calculating the average accuracy without considering MRA, we divide by six, using only the accuracy score from VSI-Bench along with the scores from the remaining datasets.

# C. Inference Time and GPU Memory

Table 5 presents the inference time, measured in seconds, of the baseline Qwen-2.5-VL-7B, `V-Reason`-7B, `V-Reason`-7B `(Lite)`, and Video-R1 across the six video reasoning benchmarks. All experiments were conducted on input videos with maximum video pixels set to $128 \times 28 \times 28$ and 32 frames temporal length. The reported results are the average over 50 samples.

From the results, it is evident that `V-Reason` and `V-Reason(Lite)` consistently outperforms Video-R1 in terms of wall-clock inference time except for VideoMMMU. Specifically, `V-Reason` reduces inference time by approximately **20–67%** compared to Video-R1 across the evaluated benchmarks. For instance, on TempCompass, the inference time decreases from 11.8 seconds per sample to 3.9 seconds per sample, while on MVBench, the reduction is from 10.7 seconds per sample to 4.1 seconds per sample. Fig. 1(d) shows that `V-Reason` has the maximum average output token count for VideoMMMU dataset and

**Algorithm 1** Autoregressive LMM inference with `V-Reason`

**Require:** Pretrained LLM $f_\theta$; Encoder $\mathcal{E}$; video frames $\mathcal{V}$; text prompt $\mathcal{X}$; Sampler $\textsc{Sample}(;)$; maximum length $L_{\max}$; temperature $\tau$; vocabulary $\mathcal{W}$.

**Ensure:** Generated text $\hat{\mathbf{y}}$.

**function** $\textsc{UpdateV}(\mathsf{V})$
  $\mathsf{V_L}' \leftarrow \frac{\mathbb{V}_L+\Delta V}{||\mathbb{V}_L+\Delta V||}||\mathbb{V}_L||$ ▷ add trainable offset and normalize
  **return** $\mathsf{V_L}'$
**end function**
**function** $\textsc{Optimize}(\boldsymbol{\ell}_\mathsf{N}, \Delta\mathsf{V}, k)$
  $p_k \leftarrow \textsc{Softmax}(\boldsymbol{\ell}_N)$
  $H_k \leftarrow -\sum_{i\in|\mathcal{W}|} p_k^i(\Delta V)\log(p_k^i(\Delta V))$
  $\alpha_k = \begin{cases} -1 \text{ if } H_k^{ema} < H_{peak}^{ema}, \\ +1 \text{ otherwise,} \end{cases}$ ▷ compute alpha
  $\mathcal{L}_{switch}(p_k; \Delta\mathsf{V}) \leftarrow -\alpha_k H_k$ ▷ compute loss
  $\Delta\mathsf{V} \leftarrow \arg\min \mathcal{L}_{switch}(p_k; \Delta\mathsf{V})$ ▷ update parameters
  **return** $\Delta\mathsf{V}$
**end function**
$\mathbf{z}_{1:N} \leftarrow \mathcal{E}(\mathcal{V}, \mathcal{X})$
$(\boldsymbol{\ell}_N, \mathsf{KV}) \leftarrow f_\theta(\mathbf{z}_{1:N})$ ▷ prefill: compute logits and full KV cache
$\hat{y}_1 \leftarrow \textsc{Sample}(\boldsymbol{\ell}_N, \tau)$
$\hat{\mathbf{y}} \leftarrow [\hat{y}_1]$
$t \leftarrow 1$
**while** $t < L_{\max}$ and $\hat{y}_t \neq$ [EOS] **do**
  $\mathsf{V} \leftarrow \textsc{UpdateV}(\mathsf{V}, \mathcal{I}_v, \pi)$
  $\Delta\mathsf{V} \leftarrow \textsc{Optimize}(\boldsymbol{\ell}_N, \Delta\mathsf{V}, k)$
  $(\boldsymbol{\ell}_{N+t}, \mathsf{KV}) \leftarrow f_\theta(\hat{y}_t \mid \mathsf{KV})$
  $\hat{y}_{t+1} \leftarrow \textsc{Sample}(\boldsymbol{\ell}_{N+t}, \tau)$
  $\hat{\mathbf{y}} \leftarrow [\hat{\mathbf{y}}; \hat{y}_{t+1}]$
  $t \leftarrow t + 1$
**end while**
**return** $\hat{\mathbf{y}}$

so using a step-size of 4 results in more number of optimization steps as compared to other datasets. This explains the anomaly observed in VideoMMMU results where the inference time is higher than Video-R1-7B. Further, comparing `V-Reason` and `V-Reason(Lite)` shows that token pruning introduces additional latency that increases the inference time marginally (+0.23 seconds) as compared to the full version without any pruning. These results highlight that `V-Reason`-7B and `V-Reason`-7B `(Lite)` achieves a significant efficiency advantage in wall-clock inference time over the RL-trained model while narrowing the gap to within 0.6% accuracy as demonstrated in Table 1.

We report the peak GPU memory usage for all models and compare `V-Reason(Lite)` with `V-Reason` to show the benefit of our pruning variant in reducing GPU memory requirements. Table 6 shows that both `V-Reason` and

**Algorithm 2** Autoregressive LMM inference with `V-Reason(Lite)`

**Require:** Pretrained LLM $f_\theta$; Encoder $\mathcal{E}$; video frames $\mathcal{V}$; text prompt $\mathcal{X}$; Sampler $\textsc{Sample}(;)$; maximum length $L_{\max}$; temperature $\tau$; pruning policy $\pi$ (e.g., keep ratio $r$ by importance).

**Ensure:** Generated text $\hat{\mathbf{y}}$.

**function** $\textsc{PruneKV}(\mathsf{KV}, \mathcal{I}_v, \pi)$
  $\mathcal{S} \leftarrow \text{Score}(\mathsf{KV}, \mathcal{I}_v)$ ▷ low L2-norm
  $\mathcal{K} \leftarrow \text{Select}(\mathcal{I}_v, \mathcal{S}, \pi)$ ▷ indices to keep among video positions
  $\mathcal{M} \leftarrow \{\text{all text positions}\} \cup \mathcal{K}$ ▷ full keep-set
  $\mathsf{KV}' \leftarrow \text{IndexSelect}(\mathsf{KV}, \mathcal{M})$ ▷ prune keys/values along sequence dimension
  **return** $\mathsf{KV}'$
**end function**
$\mathbf{z}_{1:N} \leftarrow \mathcal{E}(\mathcal{V}, \mathcal{X})$
$(\boldsymbol{\ell}_N, \mathsf{KV}) \leftarrow f_\theta(\mathbf{z}_{1:N})$ ▷ prefill: compute logits and full KV cache
$\mathcal{I}_v \leftarrow \{1, \ldots, N_v\}$ ▷ positions of video tokens
$\mathsf{KV} \leftarrow \textsc{PruneKV}(\mathsf{KV}, \mathcal{I}_v, \pi)$ ▷ KV-cache pruning for efficiency
$\hat{y}_1 \leftarrow \textsc{Sample}(\boldsymbol{\ell}_N, \tau)$
$\hat{\mathbf{y}} \leftarrow [\hat{y}_1]$
$t \leftarrow 1$
$\hat{\mathbf{y}} \leftarrow \text{AutoRegressive}[\hat{\mathbf{y}}; \hat{y}_1]$ ▷ inference optimization same as **Algorithm 1**
**return** $\hat{\mathbf{y}}$

`V-Reason(Lite)` increase the memory overhead slightly compared to the baseline Qwen-2.5-VL-7B and the Video-R1-7B model as expected due to the additional memory overhead in optimization. Note that the memory overhead is much lower than optimizing for all decoder layers in the KV-cache. To further reduce the overhead, we introduced the lite variant `V-Reason(Lite)`. The table shows that `V-Reason(Lite)` reduces the average memory requirement across all datasets by **11.6%** as compared to the full variant. In particular, the memory requirements drop by **20%** on datasets with longer output token count length such as VideoMMMU (see Fig. 1(d)) suggesting the effectiveness of the proposed Lite variant. Notably, the peak GPU memory of `V-Reason(Lite)` method is always lower than 32GB for the 7B model (on the datasets tested). This shows that the proposed lite variant is more suited for relatively smaller GPUs (e.g., 32GB V100 GPUs) and would not require more expensive GPUs like the Full variant.

**Trainable memory computation example for the controller.** Let us assume a fixed video token length of 1920 for analysis. Then the proposed controller introduces a parameter tensor of shape $(1, 4, 1920, 128)$ for Qwen-2.5-VL-7B model, amounting to $N = 983{,}040$ trainable scalars.

Table 5. Inference time (in seconds/sample) of Qwen-2.5-VL-7B, `V-Reason-7B (Lite)`, `V-Reason-7B`, and Video-R1-7B across different video reasoning benchmarks. Averaged over 50 samples from each dataset.

| Model | VSI-Bench | VideoMMMU | MMVU | TempCompass | MVBench | VideoMME | Average |
|---|---|---|---|---|---|---|---|
| Qwen-2.5-VL-7B | 3.80 | 9.02 | 6.73 | 2.86 | 3.30 | 4.37 | 5.01 |
| Video-R1-7B | 10.17 | **11.72** | 11.61 | 11.77 | 10.69 | 11.42 | 11.23 |
| `V-Reason-7B (Lite)` | **5.43** [↓46.6%] | 14.18 [↑21.0%] | **8.86** [↓23.7%] | **4.18** [↓64.5%] | **4.45** [↓58.4%] | **6.64** [↓41.9%] | **7.29** [↓35.1%] |
| `V-Reason-7B` | **5.06** [↓50.2%] | 13.83 [↑18.0%] | **9.28** [↓20.0%] | **3.87** [↓67.1%] | **4.13** [↓61.4%] | **6.18** [↓45.9%] | **7.06** [↓37.1%] |

Table 6. Peak GPU memory (in GB) of Qwen-2.5-VL-7B, `V-Reason-7B (Lite)`, `V-Reason-7B`, and Video-R1-7B across different video reasoning benchmarks. Averaged over 50 samples from each dataset.

| Model | VSI-Bench | VideoMMMU | MMVU | TempCompass | MVBench | VideoMME | Average |
|---|---|---|---|---|---|---|---|
| Qwen-2.5-VL-7B | 16.55 | 16.65 | 16.60 | 16.47 | 16.51 | 16.53 | 16.55 |
| Video-R1-7B | 16.70 | 16.74 | 16.73 | 16.68 | 16.70 | 16.69 | 16.71 |
| `V-Reason-7B` | 23.95 | 38.48 | 29.91 | 22.32 | 23.28 | 25.56 | 27.25 |
| `V-Reason-7B (Lite)` | **22.41** [↓6.4%] | **30.79** [↓20.0%] | **25.05** [↓16.2%] | **21.45** [↓3.9%] | **21.86** [↓6.1%] | **22.89** [↓10.5%] | **24.08** [↓11.6%] |

In FP32, this corresponds to $N \times 4$ bytes $= 3.84$ MiB of weights, while in FP16 the footprint is $1.92$ MiB. During training with the AdamW optimizer, additional memory is required for the gradient and two moment estimates of the same size as the parameters. Thus, in pure FP32 training the memory becomes $4 \times 3.75 = 15.36$ MiB (weights + gradients + $m$ + $v$). Since, the controller is used only as an additive bias (element-wise addition) the arithmetic cost is negligible ($\sim N$ adds, i.e., $< 10^6$ adds). The operation above is tiny compared to the bulk of transformer computation (attention and large dense projections), which typically entail orders of magnitude more FLOPs per token for typical hidden sizes and sequence lengths; therefore the controller's compute overhead is minimal in most deployments. Note that the total GPU memory required for inference-time optimization will also include the memory required for storing the activations and gradients of the last decoder layer in the model as discussed above.

# D. Analysis on Video duration

We investigate the effect of video duration on the performance of `V-Reason` using the VideoMME dataset, which provides annotations for short, medium, and long videos. Specifically, short videos are less than two minutes in duration, medium videos range from 4 to 15 minutes, and long videos span 30 to 60 minutes. Table 7 presents a detailed breakdown of the results for both `V-Reason` and its Lite variant across these duration categories. The full `V-Reason` model consistently achieves notable gains, with a substantial improvement on short videos (**+1.8%**) and notable gains on medium (**+0.8%**) and long (**+0.9%**) videos. The Lite variant of `V-Reason` also yields a significant improvement on short videos (**+1.8%**), comparable to the full model, but its performance decreases for medium and long videos. We attribute this decline to pruning, which likely removes important temporal or contextual details, thereby reducing accuracy for longer content.

# E. Image Understanding Results

We present image understanding results in Table 9. Our method consistently improves over Qwen2.5-VL-7B, achieving an average gain of **+0.6%**, with improvements exceeding **+1%** on TextVQA and MMVet. Although these gains are modest, this is expected given the non-reasoning centric nature of these benchmarks and the strong performance of the baseline model (highly saturated).

From the perspective of our framework, these results suggest that the benefits of inference-time optimization are not limited to video reasoning, but extend to general visual reasoning tasks as well. In particular, the improvements indicate that better control of the model's exploration–exploitation behavior at inference can elicit more reliable reasoning even when the underlying knowledge is already present. Importantly, these gains are achieved without any RL or additional training, reinforcing the effectiveness of our training-free approach as a lightweight and broadly applicable enhancement to existing multimodal models.

# F. Results on SFT/RL trained models

Our objective is **not** to replace reinforcement learning (RL), but to introduce a *training-free, inference-time optimization framework* that improves reasoning efficiency and accuracy by better exploiting the pretrained model's latent capabilities. This perspective is consistent with the view of reasoning as a search process over existing knowledge, where improved control at inference can elicit stronger reasoning behavior without additional supervision. Importantly, we do not claim to fundamentally expand the model's reasoning capacity, but rather to surface and refine it through more effective exploration–exploitation dynamics.

As shown in Table 10, `V-Reason` consistently improves both CoT-SFT and RL-trained models. The gains are naturally smaller for RL models, which already exhibit more optimal reasoning behavior, but remain consistent in the range of **+0.2-0.7%**. This aligns with our hypothesis that RL primarily improves the search process, which our method approximates at inference time through controlled optimiza-

Table 7. Comparison of Qwen-2.5-VL-7B, `V-Reason`, and `V-Reason-7B (Lite)` on VideoMME dataset. The differences with the baseline are denoted in red and green colors.

| Model | Mean Acc. | Short | Medium | Long |
|---|---|---|---|---|
| Qwen-2.5-VL-7B | 53.7 | 64.6 | 50.4 | 46.1 |
| `V-Reason-7B (Lite)` | 53.5 [–0.2] | **66.4 [+1.8]** | 49.7 [–0.8] | 44.3 [–1.8] |
| `V-Reason-7B` | **54.9 [+1.2]** | **66.4 [+1.8]** | **51.2 [+0.8]** | **47.0 [+0.9]** |

Table 8. Ablation studies on pruning and learning rates for the variant using 128 px and 32 frames.

| **Model** | **VSI-Bench** (Acc/MRA) | **VideoMMMU** | **MMVU** (mc) | **MVBench** | **TempCompass** | **VideoMME** (wo sub) | Avg | Avg (wo mra) |
|---|---|---|---|---|---|---|---|---|
| Qwen2.5-VL-7B | 28.1 (33.8/22.3) | 45.8 | 61.3 | 60.7 | 72.4 | 53.7 | 50.0 | 54.6 |
| Qwen2.5-VL-7B + 50% Pruning | **28.2 (34.5/21.9) [+0.1]** | **45.1 [–0.7]** | **61.3 [+0.0]** | **60.0 [–0.7]** | **72.8 [+0.4]** | **52.8 [–0.9]** | **49.8 [–0.2]** | **54.4 [–0.2]** |
| `V-Reason-7B (Lite)`; (lr: 3e-4) | **30.5 (37.3/23.7) [+2.4]** | **46.7 [+1.6]** | **64.8 [+3.7]** | **60.6 [–0.1]** | **72.3 [–0.1]** | **53.5 [–0.2]** | **51.3 [+1.3]** | **55.9 [+1.3]** |

tion. In contrast, larger gains are observed for CoT-SFT models (up to **+0.9%**), where reasoning behavior is less optimized and thus benefits more from improved inference-time control.

Notably, these improvements are achieved without any additional training, supervision, or modification of model parameters, highlighting the effectiveness of entropy-guided inference-time adaptation. Furthermore, our method remains complementary to existing training paradigms, suggesting that it can be seamlessly integrated with both SFT and RL-based approaches. We also highlight the substantial **+5.1%** gain on Qwen-3-VL-8B (Table 2 main paper), demonstrating that inference-time optimization can yield significant improvements even for strong pretrained models.

## G. Ablation Studies

In this section, we present additional ablation studies to assess the impact of hyperparameters used during inference-time optimization, including optimization step-size (update frequency), and we further analyze the frequency of alpha values before entropy maxima.

**Pruning-Only.** Table 8 compares `V-Reason` to a baseline model that implements pruning only. This shows that it is effective in maintaining the original performance with only **-0.2%** decrease on average across all datasets. Surprisingly, it also has small gains over the baseline on the VSI-Bench and TempCompass datasets. When `V-Reason` is combined with pruning, the average gain (without MRA) increases from $-0.2$ to **1.3**. This shows that the reasoning gains derive mostly from the inference optimization.

**Learning rate.** Table 8 also reports results for `V-Reason` with a fixed learning rate of `3e-4` across six datasets. The method maintains the average performance reported in Table 1 under this setting with similar gains observed on VSI-Bench, VideoMMMU, and MMVU datasets and only a negligible drop in the performance on MVBench, Tempcompass, and VideoMME datasets, highlighting its robustness to variations in optimization hyperparameters.

**Optimization Step-size.** Figure 4 shows an ablation on optimization step-size on MMVU dataset. It shows that accuracy increases with decreasing step-size. Since smaller step-sizes correspond to more optimization steps, there is a trade-off between efficiency and accuracy (fewer steps lead to faster inference). Notably, `V-Reason` outperforms the base model for all step-sizes, demonstrating that even a few optimization steps can guide the model towards improved reasoning paths.

**Optimization Objective.** Table 3 in the paper showed that the proposed entropy-switching objective is better than alternatives such as min-entropy or max-entropy losses. Fig. 7 now shows how the alternative optimization objectives fail to reproduce the characteristic entropy dynamics (later peak, lower final entropy) associated with better reasoning, supporting the hypothesis that structured entropy regulation is key.

**Alpha Switching.** Figure 5 shows the histogram of alpha values before the EMA peak is attained. `V-Reason` sacrifices a few micro-exploration steps ($\alpha = 1$) for a substantially larger number of micro-exploitation steps ($\alpha = -1$), suggesting that it pursues more alternative paths during macro-exploration. This lengthens the macro exploration stage and delays the overall entropy peak.

## H. Entropy Curves with Other Models

**Entropy Analysis of Llama Models.** We extend our analysis to the Llama family in Figure 6, examining Llama 3.1-8B-Instruct and Llama 3.2-3B-Instruct. It shows the entropy curves averaged over a subset of 100 samples on the MATH dataset. The observed entropy trajectories demonstrate macro-exploration and macro-exploitation phases that include the micro-exploration and micro-exploitation cycles

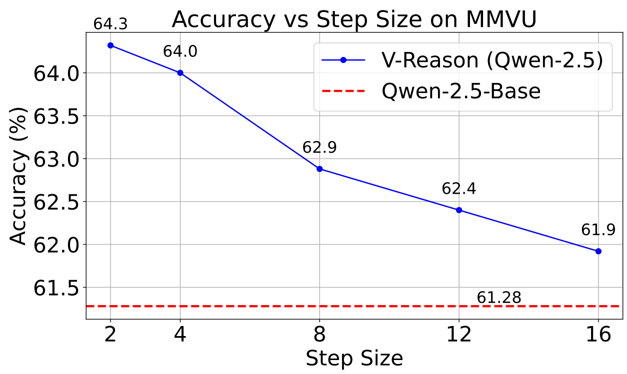



Figure 4. Optimization step-size ablations.

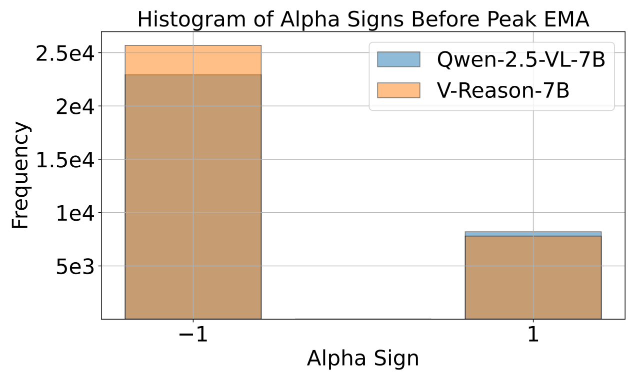
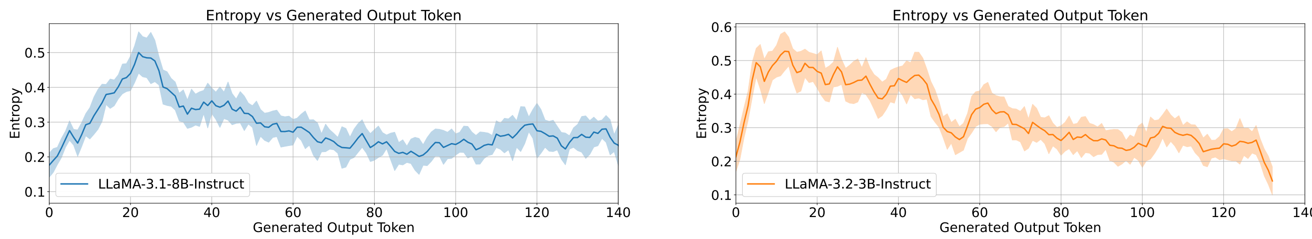



Figure 5. Alpha histogram before peak EMA entropy.




(a) Llama 3.1

(b) Llama 3.2

Figure 6. *Llama analysis*: Entropy of the output distribution averaged over 100 samples of the MATH dataset [2]. Similar to Qwen LMMs, we see clear macro-exploration and macro-exploitation phases (having micro-exploration and micro-exploitations). Note that the entropies cannot be directly compared due to different training recipes and tokenizers across the different models but we do see that the larger model (more accurate) has a delayed entropy peak than the smaller model consistent with the trends seen in Qwen model series.

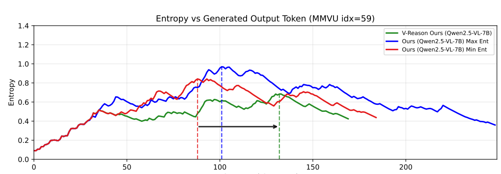



Figure 7. Comparison of different optimization objectives. The arrow denotes the shift in the peak, i.e., longer exploration for `V-Reason`.

as described in the paper. Figure 6 shows that our insights are not restricted for Qwen-based architectures. Consistent with our prior observations, the larger, higher-performing model exhibits a delayed entropy maximum compared to the smaller model, suggesting that our approach generalizes across diverse LLM architectures. This consistency across model families provides evidence that our entropy-based characterization extends beyond a single architectural lineage. We plot the models separately as the absolute entropy values are not directly comparable owing to differences in pretraining procedures and tokenization schemes.

**Entropy Analysis of Phi Models.** Continuing our cross-architecture analysis, we examine Phi 3-mini-4k-Instruct in Figure 8 by averaging over a subset of 100 samples on the MATH dataset. The generated output token entropy behavior follows the same trend observed in both Qwen and Llama families with the macro-exploration/exploitation structure, providing additional evidence for the generality of these patterns.

**Entropy Analysis of Qwen LLMs.** Figure 8 shows the entropy curves of Qwen2.5 based LLMs averaged over a subset of 100 samples on the MATH dataset. It shows that the larger 7B model (more accurate) shows a delayed peak and lower entropy overall. These trends are consistent to those observed in the video models suggesting that the proposed approach can potentially be extended to LLMs. We leave this for future work since direct application of our value-cache controller is not feasible with text-only inputs as they lack the distinct modality embeddings (video tokens) that provide optimization targets in multimodal models. Addressing this requires non-trivial contributions which is beyond the scope of the current work. However, it is very exciting to see similar macro-exploration and macro-exploitation trends with cycles of micro-exploration and micro-exploitation that cause a delayed entropy maximum for the better model across architectures and even for LLMs.

## I. Limitations

Although `V-Reason` demonstrates consistent improvements across benchmarks, there are certain limitations. First, our approach relies on the knowledge of the pretrained model to explore alternative paths during the thinking process and so for certain tasks that are under-represented in the pre-trained model, `V-Reason` can only provide modest gains compared to training-based approaches. Second, the Lite variant, while improving memory efficiency, incurs a measurable drop in accuracy for medium and long-duration videos,

Table 9. Image Understanding benchmarks (Qwen2.5-VL,V-Reason)

| Method | GQA | TextVQA | SQA | SEED | MMVet | Avg. |
|---|---|---|---|---|---|---|
| Qwen2.5-VL-7B | 48.9 | 66.4 | **42.5** | **74.1** | 45.8 | 55.5 |
| V-Reason-7B (Lite) | **49.2** [+0.2] | **67.7** [+1.3] | **42.8** | **74.1** | **46.9** [+1.1] | **56.1**[+0.6] |

Table 10. Comparison of performance of V-Reason on CoT-SFT/RL trained models.

| Model | VSI-Bench | V.MMMU | MMVU | MVBench | TCompass | V.MME | Avg | Avg wo mra |
|---|---|---|---|---|---|---|---|---|
| COT-SFT-7B | 32.5 (**32.2**/32.8) | 43.7 | 59.2 | 60.9 | 69.5 | 54.9 | 50.4 | 53.4 |
| +V-Reason | **32.6** (32.1/**33.1**) | **45.1**[+1.4] | **61.8**[+2.6] | **62.0**[+1.1] | **69.6**[+0.1] | **55.1**[+0.2] | **51.3**[+0.9] | **54.3**[+0.9] |
| Video-R1-7B | **35.6** (**30.9**/39.2) | **48.8** | 64.0 | 64.1 | **73.3** | 58.7 | **54.1** | 56.6 |
| +V-Reason | 34.2 (29.1/**39.3**) | **48.8** | **64.2**[+0.2] | **64.8**[+0.7] | **73.3** | **59.4**[+0.7] | **54.1** | **56.7**[+0.1] |

suggesting that pruning may discard valuable temporal information for those cases. Such limitations can be investigated as future work, as described next.

## J. Future Work

To our knowledge, V-Reason is the first work that targets the *video reasoning without training* problem. Hence, a number of exciting avenues exist for future research.

First, our entropy-based objective is applied only at inference time; integrating it into model training could potentially yield stronger gains and is an avenue for potential future research. Other directions of future research include exploring alternative inference-time metrics and loss functions that can further enhance reasoning.

Second, as a training-free framework, our method does not leverage task-specific supervision, which may limit its ability to capture nuanced reasoning strategies compared to reinforcement learning-based approaches. Hence, a combination of supervised finetuning and inference-time optimization-based reasoning techniques can also be explored in the future. Additionally, tailored solutions that can handle longer videos for the Lite variant can also be investigated.

Finally, although our proposed approach is motivated for videos, the idea of entropy-based inference-time optimization for enhanced reasoning is generic and can be extended to large language models (LLMs). We conducted a preliminary analysis of the entropy behavior of language models for MATH reasoning tasks and observed similar trends as the video models as discussed in section H.

## K. Additional Qualitative results

Figures 9, 10, 11, 12, 13, and 14 shows additional examples where the baseline Qwen-2.5-VL-7B failed to arrive at the correct solution while V-Reason arrived at the correct answer following an alternative reasoning path. The examples span a variety of topics in science, engineering, action recognition and counting. In particular, Figure 13 shows an example where the base model becomes distracted by irrelevant context while V-Reason correctly interprets the frames. Similarly, for the counting task in Figure 14, the base model provides confident yet incorrect totals, while V-Reason avoids these mistakes and arrives at the correct answer through effective comparison.

Figures 15, 16, 17, and 18 show examples where both the baseline Qwen-2.5-VL-7B and V-Reason arrive at the correct answer while going through similar or alternative reasoning traces. In all these examples, V-Reason shows the consistent trend of longer exploration (delayed peak) and lower overall entropy induced by the micro-exploration and micro-exploitation cycles in our proposed optimization objective. In particular, Figure 15 shows that V-Reason arrives at the correct solution using fewer output tokens as compared to the baseline as illustrated in Figure 1(d). Interestingly, in Figure 17 V-Reason uses more output tokens to provide the correct answer as compared to the baseline. However, V-Reason results in a more confident answer as seen from the lower overall entropy as compared to the baseline. This trend of higher confidence and lower overall entropy is seen in all the examples of V-Reason suggesting the effectiveness of our macro-exploitation phase induced by our proposed objective function.

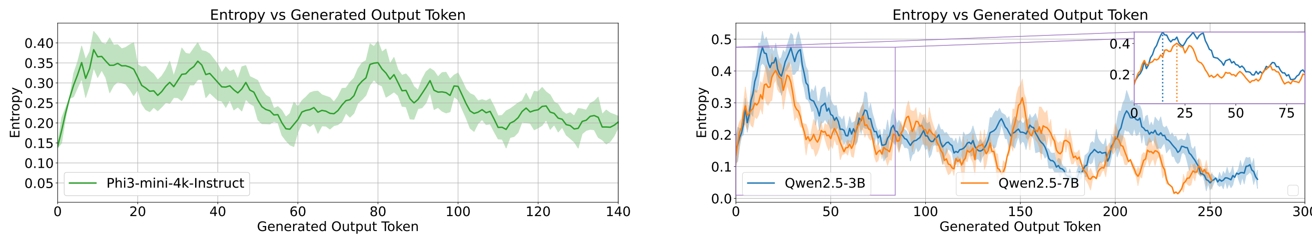
## L. LLM Usage

LLM was used to polish the writing (e.g., grammar, clarity). It was not used for literature retrieval, discovery of related work, or research ideation.

## References


[1] Erfan Baghaei Potraghloo, Seyedarmin Azizi, Souvik Kundu, and Massoud Pedram. Top-h decoding: Adapting the creativity and coherence with bounded entropy in text generation. *NeurIPS*, 2025. submitted / available on arXiv, code: https://github.com/ErfanBaghaei/Top-H-Decoding. 2

[2] An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, Dayiheng Liu, Jianhong Tu, Jingren Zhou, Junyang Lin, et al. Qwen2. 5-math technical report: Toward mathematical expert model via self-improvement. *arXiv preprint arXiv:2409.12122*, 2024. 6, 8




(a) Phi 3

(b) Qwen 2.5

Figure 8. *Phi and Qwen analysis*: Entropy of the output distribution averaged over 100 samples of the MATH dataset [2]. (a) Similar to Qwen LMMs, we observe clear macro-exploration and macro-exploitation phases (with micro-exploration and micro-exploitation) with the Phi model. (b) Qwen LLM family also shows the similar pattern observed in the LMMs with the better model showing delayed entropy maxima and lower entropy peak.


**<u>Question:</u> Which of the following is the most accurate explanation for why the indicator on the aluminum rod melted first? A. Aluminum has a higher thermal conductivity than copper, allowing heat to travel faster along the rod. B. Aluminum has a lower heat capacity per unit volume than copper, so it heats up faster despite having a lower thermal conductivity. C. The melting point of aluminum is lower than that of copper, causing the indicator to melt sooner. D. Aluminum rods have a larger cross-sectional area, reducing thermal resistance. E. The experiment demonstrates experimental error; copper should have heated up faster in theory.**

**<u>Qwen-2.5-VL-7B:</u>**

Let me think through this step by step. The question is asking why the indicator on the aluminum rod melted first. \n\nFirst, let's consider the options provided:\n- A. This suggests that aluminum's higher thermal conductivity allows heat to travel faster, which doesn't directly explain why the indicator melted first.\n- B. This option talks about heat capacity and thermal conductivity, which are important but doesn't directly address why the indicator melted first.\n- C. This option mentions the melting point of aluminum being lower than copper, which is a key factor. If aluminum melts at a lower temperature, it would indeed melt first.\n- D. This option talks about cross-sectional area, which is not directly related to the melting of the indicator.\n- E. This suggests that the experiment might have errors, but the question is asking for the most accurate explanation, not about the validity of the experiment.\n\nGiven the context of the experiment, the most accurate explanation is related to the melting point of the materials. Since aluminum has a lower melting point than copper, the indicator on the aluminum rod would melt first.\n\nTherefore, the correct answer is:\n\n<answer>C</answer>

**<u>V-Reason-7B:</u>**

Let me think through this step by step. The question is asking why the indicator on the aluminum rod melted first. The key points to consider are the thermal conductivity, heat capacity, and melting point of aluminum and copper.\n\n1. **Thermal Conductivity**: Aluminum has a higher thermal conductivity than copper. This means heat travels faster through aluminum, but this doesn't directly explain why the indicator melted first. It's more about how quickly the heat is distributed along the rod.\n\n2. **Heat Capacity**: Heat capacity per unit volume is a factor here. Aluminum has a lower heat capacity per unit volume compared to copper. This means that for a given amount of heat, aluminum will heat up faster than copper. This is because it takes less energy to raise the temperature of aluminum compared to copper.\n\n3. **Melting Point**: The melting point of aluminum is lower than that of copper. However, this is not the primary reason for the indicator melting first. The melting point is a property of the material and doesn't directly relate to the rate at which the indicator melts.\n\n4. **Cross-sectional Area**: The cross-sectional area of the rods is not mentioned, so we can't assume it's a factor here.\n\n5. **Experimental Error**: The experiment is designed to measure the specific heat of metals, and the observed behavior is consistent with the properties of the materials.\n\nGiven these points, the most accurate explanation is that aluminum has a lower heat capacity per unit volume than copper, so it heats up faster despite having a lower thermal conductivity.\n\n<answer>B</answer>

Figure 9. Qualitative result: An example output and comparison with the baseline Qwen-2.5-VL-7B. Black arrow denotes the shift in the peak demonstrating longer exploration for `V-Reason`-7B.


**Question:** **What mechanical components are needed to achieve the actions in the video? A. recirculating-ball B. rack-and-pinion C. worm-and-roller D. planetary gear set E. A, B, and C**

**Qwen-2.5-VL-7B:**

Let me think about this. The video shows a person driving a car, and the steering wheel is being turned. The options provided are related to different types of steering systems. \n\n- A recirculating-ball system is a type of steering gear that uses a ball nut and a ball screw to convert the rotary motion of the steering wheel into the linear motion of the steering shaft.\n- A rack-and-pinion system uses a rack and a pinion gear to convert the rotary motion of the steering wheel into the linear motion of the steering shaft.\n- A worm-and-roller system is another type of steering gear that uses a worm gear and a roller to convert the rotary motion of the steering wheel into the linear motion of the steering shaft.\n\nGiven that the video shows a person driving a car and the steering wheel is being turned, it is likely that the car is equipped with a rack-and-pinion steering system. This is the most common type of steering system used in modern cars.\n\n<answer>B</answer>

**V-Reason-7B:**

Let me think about this. The video shows a person driving a car, and the steering wheel is being turned. The steering system in a car can be either a recirculating-ball, a rack-and-pinion, or a worm-and-roller. The recirculating-ball and rack-and-pinion systems are more common in modern cars, while the worm-and-roller system is less common. The recirculating-ball system is known for its smooth operation and is often used in high-performance cars. The rack-and-pinion system is more common in everyday cars and is known for its efficiency and ease of use. The worm-and-roller system is less common and is more complex.\n\nGiven the context of the video, it seems to be a high-performance car, which would likely use a recirculating-ball system. However, without more specific information, it's difficult to be certain. Let's break it down: the recirculating-ball system is known for its smooth operation and is often used in high-performance cars, which matches the context of the video. The rack-and-pinion system is more common in everyday cars and is known for its efficiency and ease of use. The worm-and-roller system is less common and is more complex.\n\n<answer>E</answer>

Figure 10. Qualitative result: An example output and comparison with the baseline Qwen-2.5-VL-7B. Black arrow denotes the shift in the peak demonstrating longer exploration for V-Reason-7B.


**Question:** **What is the explanation of the shown phenomenon? A. Because the plastic rod is charged, after direct contact with the iron pieces, the iron pieces acquire opposite charges, causing them to attract each other. B. Because the plastic rod is charged and is not in direct contact with the iron pieces, the iron pieces are inductively charged and carry the same charge and repel each other. C. Because the air ambient temperature is high, the iron sheets expand and separate when heated. D. Because the plastic rod has a high temperature, the heat is transferred to the iron sheets causing them to expand and separate. E. Because the plastic rod is magnetized, and when brought near the iron pieces, induces magnetic forces causing the iron pieces to attract each other.**

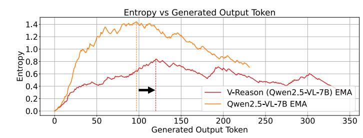
**Qwen-2.5-VL-7B:**

Let me think about this. The video shows a plastic rod being rubbed with a cloth and then brought near a set of iron pieces inside a bell jar. The iron pieces are observed to move, indicating a force of attraction or repulsion. The text in the video suggests that the rod is charged and not in direct contact with the iron pieces, which aligns with the concept of electrostatic induction.\n\nThe phenomenon described in the video is consistent with the principle of electrostatic induction. When a charged object (in this case, the plastic rod) is brought near uncharged objects (the iron pieces), it induces a charge on the iron pieces. If the rod is positively charged, it will induce a negative charge on the iron pieces that are closer to it, and a positive charge on the iron pieces that are farther away. This causes the iron pieces to repel each other.\n\nGiven this reasoning, the correct explanation is:\n\nB

**V-Reason-7B:**

Let me think about this. The video shows a plastic rod being brought near iron pieces without touching them, and the iron pieces move apart. This phenomenon is not due to temperature or magnetism, as the rod is not hot and the iron pieces do not show magnetic attraction. The rod is not in direct contact with the iron pieces, so it cannot transfer heat or magnetism. The iron pieces are not expanding due to heat, as they are not in a high-temperature environment. The only logical explanation is that the rod is charged and induces a charge in the iron pieces, causing them to repel each other.\n\n<answer>B</answer>

Figure 11. Qualitative result: An example output and comparison with the baseline Qwen-2.5-VL-7B. Black arrow denotes the shift in the peak demonstrating longer exploration for V-Reason-7B.




**Question: Which of the following best explains the phenomenon on the beaker in the video? A. The internal air pressure increases to about 1.05 atm due to thermal expansion; this pressure increase forces warm air out of the object, cooling the surrounding air and causing condensation on the beaker. B. The internal air pressure decreases to about 0.90 atm due to thermal contraction; this pressure drop causes air to flow into the object, warming the surrounding air and preventing condensation on the beaker. C. The internal air pressure remains at 1 atm because the temperature change is insufficient to affect the pressure significantly; condensation occurs due to humidity changes in the room. D. The internal air pressure decreases to about 0.94 atm due to thermal contraction; the pressure difference draws moist ambient air toward the cold surface of the beaker, where it cools below the dew point and condenses. E. The internal air pressure decreases slightly, but the condensation is primarily due to evaporation of the cold water cooling the beaker surface below the ambient dew point.**

**Qwen-2.5-VL-7B:**

Let me think through this step by step. The video shows a beaker with a thermometer inside, and a person is manipulating a red object, possibly a hot object, near the beaker. The beaker is filled with water, and the thermometer is submerged in it. The red object is being moved in and out of the water, and the water level in the beaker changes slightly.\n\nThe phenomenon observed is likely related to the temperature change of the water and the resulting pressure changes. When the red object is moved into the water, it likely causes a temperature change in the water, which in turn affects the pressure inside the beaker.\n\nLet's break it down:\n- The red object is likely heated, causing the water to warm up.\n- As the water warms up, it expands, increasing the internal air pressure in the beaker.\n- This increased pressure forces the warm air out of the beaker, cooling the surrounding air and causing condensation on the beaker.\n\nGiven this reasoning, the best explanation is that the internal air pressure increases due to thermal expansion, forcing warm air out and causing condensation.\n\n<answer>A</answer>

**V-Reason-7B:**

Let me think through this step by step. The video shows a beaker with a condensation phenomenon occurring on its surface. Condensation typically happens when the temperature of the beaker is lower than the surrounding air, causing the moisture in the air to condense on the beaker's surface. This suggests a temperature difference between the beaker and the surrounding air.\n\nNow, let's consider the options:\n\nA. This option talks about thermal expansion and condensation due to air pressure changes. However, the video doesn't show any significant change in the beaker's temperature or pressure that would cause such a phenomenon.\n\nB. This option suggests a decrease in internal air pressure due to thermal contraction, which would cause air to flow into the object, warming the surrounding air and preventing condensation. This doesn't align with the observed condensation.\n\nC. This option states that the internal air pressure remains at 1 atm, which is unlikely given the temperature change. Also, it doesn't explain the condensation.\n\nD. This option explains that the internal air pressure decreases due to thermal contraction, drawing moist ambient air toward the cold surface of the beaker, where it cools below the dew point and condenses. This aligns well with the observed condensation.\n\nE. This option suggests that the condensation is primarily due to evaporation of the cold water cooling the beaker surface, which doesn't explain the condensation on the beaker's surface.\n\nAfter considering all the options, the most logical explanation is that the internal air pressure decreases due to thermal contraction, drawing moist ambient air toward the cold surface of the beaker, where it cools below the dew point and condenses.\n\n<answer>D</answer>

Figure 12. Qualitative result: An example output and comparison with the baseline Qwen-2.5-VL-7B. Black arrow denotes the shift in the peak demonstrating longer exploration for V-Reason-7B.


**Question**: What is special about the celebration in New York according to the video? A. Hosting large parades. B. Dressing in green and dyeing the river to green. C. Drinking a lot. D. Planting shamrocks.

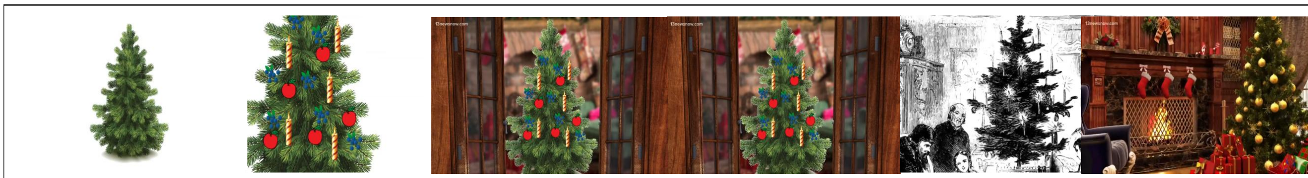
Qwen-2.5-VL-7B:

Let me think about this. The video mentions that 2 million spectators watch the NYC parade, which indicates a large turnout. However, it also specifically states \"40 LBS OF DYE USED TO CHICAGO RIVER GREEN\" and \"2 MILLION SPECTATORS WATCH NYC PARADE\". This suggests that the NYC parade is notable for its size and the fact that the river is dyed green. \n\nNow, let's break it down. The question asks what is special about the celebration in New York. The options provided are:\nA. Hosting large parades.\nB. Dressing in green and dyeing the river to green.\nC. Drinking a lot.\nD. Planting shamrocks.\n\nThe video clearly highlights the river being dyed green, which is a unique and special feature of the NYC St. Patrick's Day celebration. \n\nTherefore, the answer is B.

V-Reason-7B:

Let me think about this. The video mentions a specific detail about New York's celebration. It states \"2 MILLION SPECTATORS WATCH NYC PARADE\". This suggests a large turnout for the parade. However, the question asks for something special about the celebration in New York. The video also mentions \"40 LBS OF DYE USED TO CHICAGO RIVER TURN THE RIVER GREEN\". This detail is about Chicago, not New York. \n\nNow, let's break it down. The video does not provide information about New York specifically dyeing a river green. It only mentions a parade and a large number of spectators. \n\nOh, I see. The video does not provide information about New York specifically dyeing a river green. It only mentions a parade and a large number of spectators. \n\nTherefore, the special aspect of the celebration in New York according to the video is hosting large parades.\n\n<answer>A</answer>

Figure 13. Qualitative result: An example output and comparison of V-Reason-7B with the baseline Qwen-2.5-VL-7B.



**Question**: When demonstrating the Germany modern Christmas tree is initially decorated with apples, candles and berries, which kind of the decoration has the largest number? A. Apples. B. Candles. C. Berries. D. The three kinds are of the same number.

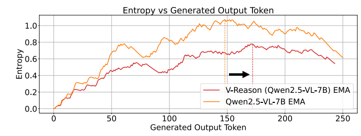
Qwen-2.5-VL-7B:

Let me think about this. The question is asking about the number of decorations on a Germany modern Christmas tree, specifically apples, candles, and berries. The image shows a Christmas tree decorated with these items. \n\n<think>\nI need to count the decorations for each type:\n- Apples: There are 3 apples visible.\n- Candles: There are 3 candles visible.\n- Berries: There are 3 berries visible.\n\nAll three types of decorations appear to be present in equal numbers.\n</think>\n\n<answer>\nD\n</answer>

V-Reason-7B:

Let me think about this. The question is asking about the number of decorations on a Germany modern Christmas tree, specifically apples, candles, and berries. The image shows a Christmas tree with these decorations. \n\n<think>\nI need to count the decorations on the tree. There are apples, candles, and berries. It's a bit difficult to count precisely without a clear image, but from what I can see, the berries seem to be the most numerous. The apples and candles are fewer in number compared to the berries.\n</think>\n\n<answer>\nC\n</answer>

Figure 14. Qualitative result: An example output and comparison of V-Reason-7B with the baseline Qwen-2.5-VL-7B.


**Question: Which of the following best explains the primary benefits brought by the treaty shown in the video? A. By creating a common market allows free movement of goods, services, capital, and labor, letting members specialize by their strengths. B. By adopting a single currency to eliminate exchange rate fluctuations and reduce transaction costs among member states. C. By imposing uniform protectionist tariffs against non-member countries to shield domestic industries from external competition. D. By integrating fiscal policies to redistribute wealth from richer to poorer member states, reducing economic disparities. E. By standardizing labor laws and working conditions across member states to prevent a 'race to the bottom' in labor standards.**

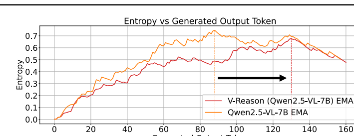
**Qwen-2.5-VL-7B:**

Let me think about this. The video seems to be discussing the Treaty of Rome, which is the founding treaty of the European Economic Community (EEC), the predecessor of the European Union (EU). The treaty aimed to create a common market and economic union among its member states. \n\nOption A talks about the creation of a common market allowing free movement of goods, services, capital, and labor, which aligns with the goal of the EEC to integrate economies and facilitate trade among member states.\n\nOption B mentions adopting a single currency, which was not part of the initial EEC treaty but became a part of the later European Union.\n\nOption C is about protectionist tariffs, which is not a primary benefit of the EEC treaty. The EEC aimed for free trade among member states.\n\nOption D is about fiscal policies and wealth redistribution, which is not a primary focus of the EEC treaty.\n\nOption E is about standardizing labor laws, which was also not a primary focus of the EEC treaty.\n\nGiven the context of the video and the options provided, the primary benefit brought by the treaty shown in the video is the creation of a common market allowing free movement of goods, services, capital, and labor.\n\n<answer>A</answer>

**V-Reason-7B:**

Let me think about this. The video is about the Treaty of Rome, which is the founding treaty of the European Economic Community (EEC), the precursor to the European Union (EU). The primary benefits of such a treaty would be related to the creation of a common market and the facilitation of trade and economic integration among member states.\n\nOption A talks about the creation of a common market that allows for the free movement of goods, services, capital, and labor. This aligns well with the goals of the EEC, which aimed to create a single market and a customs union.\n\nOption B mentions adopting a single currency. While the EU later adopted the euro, this was not part of the original Treaty of Rome.\n\nOption C is about protectionist tariffs, which is the opposite of what the EEC aimed for, as it sought to create a free trade area.\n\nOption D is about fiscal policies and wealth redistribution, which is not a primary benefit of the EEC.\n\nOption E is about standardizing labor laws, which is also not a primary benefit of the EEC.\n\nGiven this analysis, the best explanation for the primary benefits brought by the treaty shown in the video is:\n\n<answer>A</answer>

Figure 15. Qualitative result: An example output and comparison with the baseline Qwen-2.5-VL-7B. Black arrow denotes the shift in the peak demonstrating longer exploration for V-Reason-7B.




**Question: Which of the following best explains the effect of increasing resistance in the circuit on the voltmeter readings over time? A. The increased resistance causes the capacitor to charge faster, so the voltmeter reading increases more rapidly. B. The increased resistance causes the capacitor to charge more slowly initially, but the final voltage reached is higher. C. The increased resistance has no effect on the charging process or the voltmeter readings since the battery voltage is constant. D. The increased resistance causes the capacitor to charge more slowly, so the voltmeter reading increases more slowly. E. The increased resistance causes the capacitor to discharge, causing the voltmeter reading to decrease over time.**

**Qwen-2.5-VL-7B:**

Let me think about this. When resistance increases in a circuit, it affects the rate at which the capacitor charges. The voltmeter measures the voltage across the capacitor. If the resistance increases, the current through the circuit decreases, which means the capacitor charges more slowly. This slower charging process would result in a slower increase in the voltage reading on the voltmeter.\n\nNow, let's break it down:\n- Increasing resistance means less current flows through the circuit.\n- Less current means the capacitor charges more slowly.\n- Slower charging means the voltage on the voltmeter increases more slowly.\n\nGiven these points, the correct answer should reflect that the increased resistance causes the capacitor to charge more slowly, leading to a slower increase in the voltmeter reading.\n\n<answer>D</answer>

**V-Reason-7B:**

Let me think about this. Increasing resistance in a circuit generally means that the current flowing through the circuit will decrease. This is because resistance is a measure of how much a material opposes the flow of electric current. When the resistance increases, the current decreases, which in turn affects the rate at which the capacitor charges.\n\nNow, let's consider the effect on the voltmeter readings. The voltmeter measures the voltage across the capacitor. When the resistance increases, the current decreases, and this means that the capacitor will charge more slowly. As a result, the voltage across the capacitor will increase more slowly as well.\n\nSo, the correct option should reflect that the increased resistance causes the capacitor to charge more slowly, leading to a slower increase in the voltmeter reading.\n\n<answer>D</answer>

Figure 16. Qualitative result: An example output and comparison with the baseline Qwen-2.5-VL-7B. Black arrow denotes the shift in the peak demonstrating longer exploration for V-Reason-7B.


**Question: Which of the following best explains the effect of increasing resistance in the circuit on the voltmeter readings over time? A. The increased resistance causes the capacitor to charge faster, so the voltmeter reading increases more rapidly. B. The increased resistance causes the capacitor to charge more slowly initially, but the final voltage reached is higher. C. The increased resistance has no effect on the charging process or the voltmeter readings since the battery voltage is constant. D. The increased resistance causes the capacitor to charge more slowly, so the voltmeter reading increases more slowly. E. The increased resistance causes the capacitor to discharge, causing the voltmeter reading to decrease over time.**

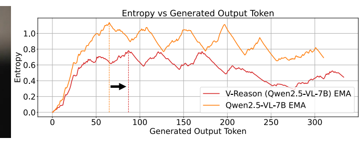
**Qwen-2.5-VL-7B:**

Let me think about this. When resistance increases in a circuit, it affects the rate at which the capacitor charges. The voltmeter measures the voltage across the capacitor. If the resistance increases, the current through the circuit decreases, which means the capacitor charges more slowly. This slower charging process would result in a slower increase in the voltage reading on the voltmeter.\n\nNow, let's break it down:\n- Increasing resistance means less current flows through the circuit.\n- Less current means the capacitor charges more slowly.\n- Slower charging means the voltage on the voltmeter increases more slowly.\n\nGiven these points, the correct answer should reflect that the increased resistance causes the capacitor to charge more slowly, leading to a slower increase in the voltmeter reading.\n\n<answer>D</answer>

**V-Reason-7B:**

Let me think about this. Increasing resistance in a circuit generally means that the current flowing through the circuit will decrease. This is because resistance is a measure of how much a material opposes the flow of electric current. When the resistance increases, the current decreases, which in turn affects the rate at which the capacitor charges.\n\nNow, let's consider the effect on the voltmeter readings. The voltmeter measures the voltage across the capacitor. When the resistance increases, the current decreases, and this means that the capacitor will charge more slowly. As a result, the voltage across the capacitor will increase more slowly as well.\n\nSo, the correct option should reflect that the increased resistance causes the capacitor to charge more slowly, leading to a slower increase in the voltmeter reading.\n\n<answer>D</answer>

Figure 17. Qualitative result: An example output and comparison with the baseline Qwen-2.5-VL-7B. Black arrow denotes the shift in the peak demonstrating longer exploration for V-Reason-7B.




**<u>Question</u>: If the solution in the video is a salt solution, what is the most likely chemical reaction equation that would occur in the test tube? A. Cu + Pb(NO3)2 → Cu(NO3)2 + Pb B. Cu + 2AgNO3 → Cu(NO3)2 + 2Ag C. Cu + FeSO4 → CuSO4 + Fe D. Cu(NO3)2 + Na2S → CuS + 2NaNO3 E. CuCl2 + Na2SO4 → CuSO4 + 2NaCl**

**<u>Qwen-2.5-VL-7B:</u>**

Let me think about this. The video shows a test tube with a wire that appears to be copper, and the solution is a salt solution. The wire seems to be reacting with the solution, and there's a visible change in the wire's surface, indicating a chemical reaction.\n\nThe options provided are all potential reactions involving copper. Let's break it down:\n\nA. This reaction involves copper reacting with lead nitrate to form copper nitrate and lead. This doesn't seem to fit the visual evidence since the wire doesn't appear to be lead.\n\nB. This reaction involves copper reacting with silver nitrate to form copper nitrate and silver. This could be a plausible reaction, as silver nitrate is a common reagent in chemistry labs and could cause a visible change in the wire's surface.\n\nC. This reaction involves copper reacting with iron sulfate to form copper sulfate and iron. This is less likely because iron is typically more reactive than copper and would likely be the one to react first.\n\nD. This reaction involves copper nitrate reacting with sodium sulfide to form copper sulfide and sodium nitrate. This is also less likely because copper nitrate is not typically used in this reaction.\n\nE. This reaction involves copper chloride reacting with sodium sulfate to form copper sulfate and sodium chloride. This is less likely because copper chloride is not typically used in this reaction.\n\nGiven the visual evidence and the options provided, the most likely reaction is the one that involves copper reacting with a silver nitrate solution, which is option B.\n\n<answer>B</answer>

**<u>V-Reason-7B:</u>**

Let me think about this. The video shows a test tube with a wire that appears to be copper, and the solution is a salt solution. The reaction seems to involve the copper wire dissolving and forming a precipitate at the bottom of the tube.\n\nOption A involves copper reacting with lead nitrate to form copper nitrate and lead. This doesn't seem to fit the observation in the video as there's no lead nitrate present.\n\nOption B involves copper reacting with silver nitrate to form copper nitrate and silver. This could be a possibility as silver nitrate is a common salt solution and could react with copper.\n\nOption C involves copper reacting with iron(II) sulfate to form copper(II) sulfate and iron. This is less likely as iron(II) sulfate is not a common salt solution and the reaction would not form a precipitate at the bottom.\n\nOption D involves copper(II) nitrate reacting with sodium sulfide to form copper sulfide and sodium nitrate. This could be a possibility as copper(II) nitrate is a common salt solution and could react with sodium sulfide to form a precipitate.\n\nOption E involves copper(II) chloride reacting with sodium sulfate to form copper(II) sulfate and sodium chloride. This is less likely as sodium sulfate is not a common salt solution and the reaction would not form a precipitate at the bottom.\n\nGiven the observation of a precipitate forming at the bottom of the tube, option B seems the most likely as it involves a common salt solution (silver nitrate) and could form a precipitate (silver) at the bottom of the tube.\n\n<answer>B</answer>

Figure 18. Qualitative result: An example output and comparison with the baseline Qwen-2.5-VL-7B. Black arrow denotes the shift in the peak demonstrating longer exploration for V-Reason-7B.