When Thinking Hurts: Mitigating Visual Forgetting in Video Reasoning via Frame Repetition

Xiaokun Sun[1,2*], Yubo Wang[1,2*], Haoyu Cao[1,2], and Linli Xu[1,2†]

[1] University of Science and Technology of China, Anhui, China
[2] State Key Laboratory of Cognitive Intelligence, Anhui, China
{sunxiaokun2020, wyb123, caohaoyu}@mail.ustc.edu.cn
linlixu@ustc.edu.cn

Abstract. Recently, Multimodal Large Language Models (MLLMs) have demonstrated significant potential in complex visual tasks through the integration of Chain-of-Thought (CoT) reasoning. However, in Video Question Answering, extended thinking processes do not consistently yield performance gains and may even lead to degradation due to "visual anchor drifting", where models increasingly rely on self-generated text, sidelining visual inputs and causing hallucinations. While existing mitigations typically introduce specific mechanisms for the model to re-attend to visual inputs during inference, these approaches often incur prohibitive training costs and suffer from poor generalizability across different architectures. To address this, we propose FrameRepeat, an automated enhancement framework which features a lightweight repeat scoring module that enables Video-LLMs to autonomously identify which frames should be reinforced. We introduce a novel training strategy, Add-One-In (AOI), that uses MLLM output probabilities to generate supervision signals representing repeat gain. This can be used to train a frame scoring network, which guides the frame repetition behavior. Experimental results across multiple models and datasets demonstrate that FrameRepeat is both effective and generalizable in strengthening important visual cues during the reasoning process.

Keywords: Video Understanding · MLLMs · Visual Reasoning

1 Introduction

Recently, multimodal large language models have demonstrated significant potential in complex visual understanding tasks by incorporating Chain-of-Thought (CoT) reasoning [10, 16, 25–27, 34, 36, 38]. However, existing studies indicate that in certain visual reasoning scenarios, as the reasoning chain grows longer, the model's attention toward visual modalities diminishes significantly [15, 21, 40].

* Equal contribution, † Corresponding author.

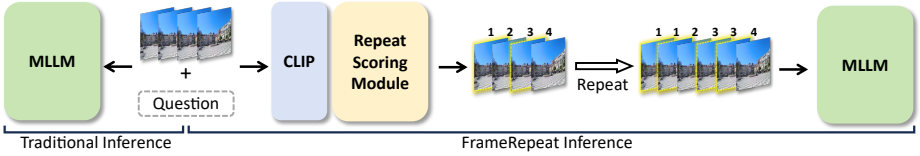

Fig. 1: Comparison between traditional inference and FrameRepeat inference. In traditional inference, video frames are uniformly sampled and directly fed into the MLLM. In contrast, the proposed FrameRepeat first scores the frames during inference, selecting the frames with the highest scores, which are then repeated at their original positions and passed into the MLLM for enhanced understanding.

The prediction process increasingly shifts toward relying on self-generated linguistic tokens, which makes the model more prone to hallucinations and reasoning drift. This phenomenon is particularly pronounced in video reasoning, where the inherent sparsity of temporal information means that critical evidence is often confined to a few sparse frames. The remaining frames, largely redundant or contextual, make it significantly harder for the model to maintain focus on salient information during reasoning. Furthermore, as videos contain much more complex information, their CoT processes are often longer than those for static images, further exacerbating this issue.

While existing approaches typically mitigate this issue by introducing specific mechanisms—such as invoking external tools to re-incorporate pivotal visual information [13, 14, 17, 44] or designing specialized loss functions to steer attention distribution [15]—they often entail computationally expensive training processes. In addition, the optimization gains achieved on a specific architecture frequently lack cross-model transferability, thereby limiting their broader application.

In contrast to the aforementioned methods, we argue that the root of this issue lies not in a deficiency of the model's reasoning capabilities, but rather in a signal imbalance between long-chain linguistic reasoning and visual input. Specifically, if the essence of the problem is the dilution of visual signals during the reasoning process, then the primary objective should not be increasing reasoning steps, but rather maintaining the effective intensity of critical visual cues amidst extended CoT sequences.

We observe a simple yet consistent phenomenon: repeating specific video frames in the input stage enhances model performance without requiring any additional training. This finding suggests that increasing the redundancy of specific frames can, to some extent, counteract the dilution effect imposed by reasoning tokens. Based on this observation, we first analyze how frame repetition influences video reasoning. By examining the shifts in attention toward repeated segments during the thinking process, alongside the variations in the log probability of correct answers, we find that repeating key frames effectively bolsters the model's focus on these frames, and the impact of repeating different frames on the output log probability varies significantly.

Accordingly, we propose elevating frame repetition from a static heuristic to a reasoning-aware visual signal re-weighting process. Specifically, we introduce FrameRepeat, an automated enhancement framework which features a lightweight repeat scoring module to predict an importance factor for each frame based on the frame's visual features and the question's semantic representation, enabling adaptive repetition of pivotal frames at the input stage of Video-LLMs. We propose Add-One-In (AOI) to train FrameRepeat: for each candidate frame, we measure the log-probability gain of the correct answer when that single frame is repeated, yielding a repeat gain signal that supervises the network to learn frame importance ranking. During inference, FrameRepeat selects the top-k frames based on the repeat scoring module's output and repeats them at their original temporal positions as shown in Fig. 1. Notably, FrameRepeat does not modify the backbone VLM architecture; instead, it strengthens the visual signal at the input level, ensuring the model maintains continuous focus on critical visual evidence throughout long-chain reasoning. Consequently, a repeat scoring module trained on one model can be seamlessly transferred to different architectures.

To comprehensively evaluate our approach, we conduct extensive experiments across two distinct backbone models, Qwen2.5-VL-Instruct-7B [2], Qwen3-VL-Thinking-8B [1] and five representative benchmarks VideoMME [11], LongVideo-Bench [35], LVBench [32], MLVU [45], Video-Holmes [6]. The results demonstrate that FrameRepeat significantly boosts model performance on video reasoning tasks while exhibiting strong generalization capabilities across different scenarios.

Our contributions can be summarized as follows:

- We analyze the phenomenon of visual signal dilution in video reasoning scenarios, where long-chain reasoning leads to a decay in visual attention and propose a novel approach that bolsters visual signal intensity through strategic frame repetition.
- We introduce FrameRepeat, a lightweight cross-attention repeat scoring module that scores each frame's importance conditioned on the question, and selects the top-scoring frames for in-place duplication at the input stage. To train this network, we propose Add-One-In (AOI), a marginal-contribution-based supervision strategy that measures each frame's value by the log-probability gain when it is individually repeated.
- Extensive experiments across five benchmarks and two distinct backbone models validate the effectiveness of FrameRepeat in enhancing video reasoning performance and demonstrate its robust generalization capabilities.

2 Related Work

2.1 Reasoning in MLLMs

Recent research has focused on enhancing the reasoning capabilities of Multimodal Large Language Models (MLLMs), inspired by the success of long chain-of-thought reasoning in language models [3, 4, 8, 20, 33, 37, 42]. LLaVA-CoT [39]

introduces a structured four-stage reasoning pipeline—summary, caption, reasoning, and conclusion—with a stage-level beam search for effective inference-time scaling. R1-Onevision [41] proposes a cross-modal formalization pipeline that converts visual inputs into formal textual representations and leverages reinforcement learning to develop generalized multimodal reasoning capabilities.

Despite these advances, recent studies have identified notable limitations in the CoT reasoning of current MLLMs. [15] find that current visual reasoning models exhibit limited visual reflection, as their attention to visual information diminishes rapidly with longer generated responses. Similarly, [40] points out that these models often excessively rely on textual information during the later stages of inference, neglecting the crucial integration of visual input. These findings highlight a critical challenge: how to prevent key visual signals from being diluted throughout the reasoning process, which is the central question explored in this work.

2.2 Video Reasoning

Extending reasoning capabilities to the video domain presents unique challenges, as it requires not only understanding individual frames but also modeling complex temporal dynamics, causal relationships, and event transitions across long sequences [5, 7, 9, 19, 22, 29–31]. Unlike image-based reasoning, video reasoning demands that models maintain a coherent understanding over time while integrating spatial and temporal cues for multi-step inference. Recent efforts have begun to explore chain-of-thought and reinforcement learning paradigms in this context. Specifically, VideoChat-R1 [18] enhances spatio-temporal perception in video reasoning through reinforcement fine-tuning. Video-R1 [10] is the first to systematically explore the R1 paradigm for video reasoning via an upgraded T-GRPO algorithm, achieving strong results on spatial-temporal benchmarks. However, in the context of video reasoning, introducing chain-of-thought often leads to performance degradation [21]. In this work, we analyze the underlying causes of this phenomenon and propose to mitigate it by enhancing the visual signal intensity of key frames.

3 Dissecting the Effect of Frame Repetition

Prior works [15, 21, 40] have found that incorporating CoT may sometimes lead to a decline in model performance. This is because, as the chain of thought grows longer, the model tends to attend more to the generated text tokens while neglecting the visual information. This phenomenon is particularly severe in video reasoning, since the effective information in a video is often concentrated in a few key frames, while the remaining visual tokens provide background or redundant information, making it even harder for the model to attend to the critical information during reasoning. Given this observation, a natural question arises: **Is there a way to highlight the key information, so that critical visual signals can maintain their effective strength throughout the reasoning process?** We conduct several experiments to investigate this.

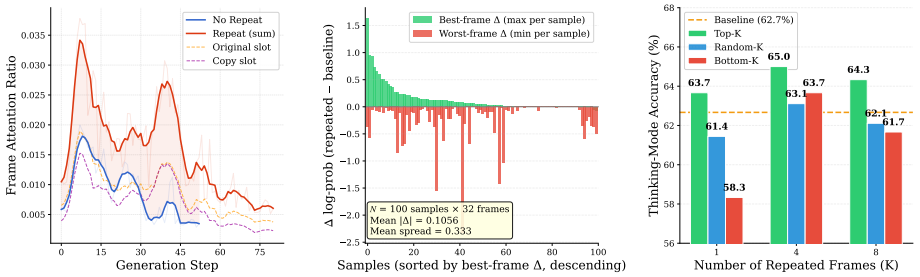

Fig. 2: Empirical analysis of frame repetition. (a) Attention ratio allocated to a repeated frame across generation steps. When a frame is repeated, it occupies two visual slots in the input sequence—the original slot and the copy slot—whose attention contributions are shown separately. Repetition significantly increases the model's overall attention to the target frame. **(b)** Per-sample best-frame (green) and worst-frame (red) Δlog-prob, showing that repetition benefit is highly frame-dependent. **(c)** Thinking-mode accuracy when repeating Top-K, Random-K, or Bottom-K frames selected by direct-mode Δlog-prob. Top-K consistently outperforms other strategies, validating the effectiveness of Δlog-prob as a frame importance signal.

3.1 Mitigating Visual Forgetting

An intuitive idea is: since critical visual signals are gradually diluted during reasoning, could we enhance the presence of these key frames in the input—for example, by repeating them—to help the model sustain attention to critical information throughout long-chain reasoning? To verify this hypothesis, we design a simple experiment to observe the effect of frame repetition on attention distribution. Specifically, we use Qwen2.5-VL as the base model and randomly sample instances from VideoMME, with each video uniformly sampled into 128 frames as input. We first perform one inference pass, identifying the key frame that receives the highest attention proportion from the answer token during answer generation, and record the attention ratio of each generated token to this key frame throughout the reasoning process. In the second pass, we duplicate this key frame and insert the copy immediately after its original position, re-running inference with 129 frames. We record the attention ratio curves over generation steps for both passes, where the attention under the repetition condition is the sum of the original and copied positions. The results are shown in Fig. 2(a). It can be clearly observed that without repetition, the model's attention to the key frame exhibits a sustained declining trend as generation progresses; with repetition, although the attention still decreases, it remains at a significantly higher level throughout the entire reasoning process.

3.2 How Frame Repetition Affects Model Predictions

To evaluate the impact of frame repetition on model outputs, we conduct a per-frame log-probability scanning experiment on Qwen2.5-VL-7B. We randomly

sample 100 instances from VideoMME, extract 128 frames per video as the baseline input, and let the model generate answers without chain-of-thought reasoning. For each sample, we repeat 32 randomly selected frames one at a time and measure the change in log-probability of the correct answer (Δ). As shown in Fig. 2(b), repeating a single frame leads to a mean $|\Delta \log\text{-prob}|$ of 0.1056, and in 94.0% of samples, at least one frame increases the probability of the correct answer. Moreover, the influence varies widely across frame positions: the average gap between the best and worst frame Δ is 0.3325 (maximum 2.3961), and in 60.0% of samples this gap exceeds 0.1.

We further investigate two questions: (1) whether the frame importance ranking obtained in direct (non-thinking) mode can guide frame selection in thinking mode, and (2) whether repeating multiple informative frames yields greater benefits than repeating a single frame. On 300 VideoMME questions with 128 base frames, we first rank 32 candidate frames by their Δlog-prob in direct mode (**Phase 1**). Then in **Phase 2**, for each $K \in \{1, 4, 8\}$, we evaluate three conditions in thinking mode: **Top-K** (largest Δ), **Bottom-K** (smallest Δ), and **Random-K** (averaged over 3 trials), along with a no-repeat baseline. As shown in Fig. 2(c), the accuracy ordering Top-K > Random-K and Random-K > Bottom-K holds consistently across all K values, confirming that direct-mode Δlog-prob is a reliable proxy for identifying frames that benefit thinking-mode reasoning. Furthermore, repeating multiple informative frames amplifies the benefit: Top-4 reaches 65.0% (+2.3% over baseline), compared to +1.0% for Top-1, indicating that the effects accumulate rather than saturate.

These findings suggest that if a method can enable the model to autonomously determine which frames to repeat, it would be possible to improve performance on video reasoning tasks without introducing additional inference steps.

4 Method

As illustrated in Fig. 3, we propose FrameRepeat, a method designed to enable Video-LLMs to autonomously learn which frames should be repeated during the input stage, thereby enhancing the signal strength of critical visual information during the reasoning process. In this section, we first introduce the architectural design of FrameRepeat. Subsequently, we present Add-One-In (AOI), a frame-level value estimation strategy based on the repeat gain of repeating one frame. This strategy measures the extent to which repeating a specific frame contributes to the model's correct response, serving as the supervision signal for training the repeat scoring module. Finally, we detail the joint training objective, which comprises both regression and ranking losses.

4.1 FrameRepeat Architecture

During our experiments, we observed that critical frames are often highly correlated with the input query. Therefore, when modeling the importance of each

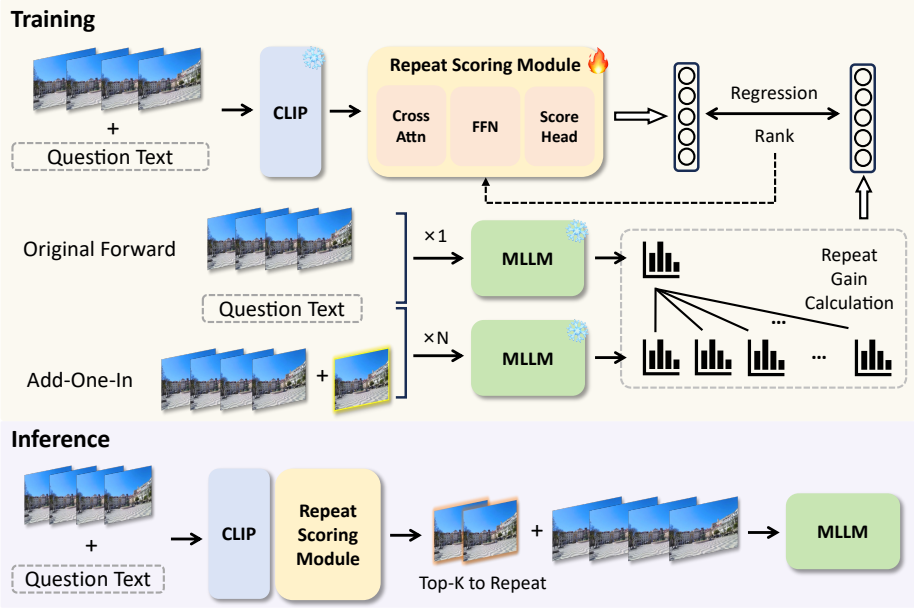

Fig. 3: Overview of the FrameRepeat framework. During training, we use the Add-One-In strategy to measure the MLLM output probability for each repeated frame. The resulting probability variations are used to compute the repeat gain, which trains the repeat scoring module. During inference, the video frames and question are passed through the scoring module to estimate the repetition importance of each frame. The top-k most important frames are then selected for repetition and fed into the MLLM for the final answer.

frame, it is necessary to incorporate textual information and enable interaction between text and visual modalities.

At the input stage, we first uniformly sample N frames from the video as in previous methods, then encode each frame and the input question through CLIP [23] to obtain their feature representations. Specifically, each frame is encoded as a d-dimensional vector $\mathbf{v}_i \in \mathbb{R}^d$ $(i = 1, \ldots, N)$, and the question text is encoded as a sequence of token embeddings $\mathbf{T} = [\mathbf{t}_1, \ldots, \mathbf{t}_L] \in \mathbb{R}^{L \times d}$, where L is the number of tokens. Since CLIP encodes each frame independently and is inherently agnostic to temporal ordering, we inject temporal information by adding sinusoidal positional encodings [28]:

$$\tilde{\mathbf{v}}_i = \mathbf{v}_i + \mathrm{PE}(i, N) \tag{1}$$

where $\mathrm{PE}(i, N)$ denotes the sinusoidal positional encoding for the i-th frame in a sequence of length N.

To enable interaction between textual and visual features, we introduce a multi-head cross-attention module. Specifically, we use the visual features as queries and the text token sequence as keys and values:

$$\mathbf{h}_i = \text{CrossAttn}(\tilde{\mathbf{v}}_i, \mathbf{T}) = \text{softmax}\left(\frac{\mathbf{W}_Q\tilde{\mathbf{v}}_i \cdot (\mathbf{W}_K\mathbf{T})^\top}{\sqrt{d_h}}\right)\mathbf{W}_V\mathbf{T} \tag{2}$$

where $\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V \in \mathbb{R}^{d\times d}$ are learnable projection matrices and d_h is the head dimension. Each frame independently attends to the full question token sequence, yielding a question-conditioned frame representation via a residual connection followed by LayerNorm:

$$\hat{\mathbf{v}}_i = \text{LayerNorm}(\tilde{\mathbf{v}}_i + \mathbf{h}_i) \tag{3}$$

After obtaining the text-fused visual features, we further refine them through a two-layer feed-forward network (FFN) with a residual connection. Since the FFN operates on features that already encode question relevance from the cross-attention layer, it is able to capture non-linear patterns in the question-frame interaction — for example, distinguishing frames that merely depict an entity mentioned in the question from those that capture the key action needed to answer it:

$$\mathbf{f}_i = \text{LayerNorm}(\hat{\mathbf{v}}_i + \text{FFN}(\hat{\mathbf{v}}_i)) \tag{4}$$

We then map each refined frame feature to a scalar importance score through a two-layer projection head:

$$\hat{s}_i = \text{ScoreHead}(\mathbf{f}_i), \quad \text{ScoreHead} : \mathbb{R}^d \to \mathbb{R}^{d/4} \to \mathbb{R}^1 \tag{5}$$

Finally, we add a zero-centered CLIP cosine similarity as a prior to the learned score:

$$s_i = \hat{s}_i + \lambda \cdot \left(\text{sim}(\mathbf{v}_i, \mathbf{t}) - \frac{1}{N}\sum_{j=1}^{N}\text{sim}(\mathbf{v}_j, \mathbf{t})\right) \tag{6}$$

where $\text{sim}(\cdot,\cdot)$ denotes the cosine similarity and λ is a scaling coefficient. This provides a stable initialization signal for training, preventing the randomly initialized network from producing arbitrary rankings. The zero-centering ensures that the prior encodes only relative relevance across frames without shifting the absolute score level. During training, the MLLM backbone and CLIP encoder remain completely frozen.

4.2 Add-One-In

The core task of the repeat scoring module is to assign an importance score to each frame such that frames with higher scores, when repeated, effectively improve the model's answer quality. This requires a training signal that directly reflects the impact of repeating a specific frame on the model's response. However, obtaining such frame-level supervision is non-trivial: existing video understanding datasets only provide question-answer pairs and lack annotations indicating which frames are critical for answering.

To address this, we propose the Add-One-In (AOI) training strategy. The key idea is to **automatically construct frame-level supervision by measuring the change in the model's log-probability of the correct answer after repeating a single frame**. Specifically, for each training sample, we first feed the original N frames into the frozen MLLM and compute the baseline log-probability of the correct answer:

$$\ell_{\text{base}} = \log P(a^* \mid \{f_1, \ldots, f_N\}, q) \tag{7}$$

where a^* is the correct answer and q is the question. Then, for each candidate frame f_i, we duplicate it once and insert the copy at its corresponding temporal position, forming an $(N+1)$-frame input, and recompute the log-probability:

$$\ell_i = \log P(a^* \mid \{f_1, \ldots, f_i, f_i, \ldots, f_N\}, q) \tag{8}$$

The difference between the two yields the *Repeat Gain* of frame f_i:

$$\Delta_i = \ell_i - \ell_{\text{base}} \tag{9}$$

A positive Δ_i indicates that repeating this frame increases the model's confidence in the correct answer, while a negative Δ_i suggests that repeating it introduces interference. This supervision signal comes directly from the MLLM's own feedback, requires no additional human annotation, and is tightly aligned with the downstream objective of answering the question correctly.

In practice, computing the repeat gain for all N frames individually incurs substantial overhead. To mitigate this, we adopt a hybrid candidate sampling strategy: we form the candidate set $\mathcal{C}$ from the union of the top-K frames ranked by the repeat scoring module's current scores and K randomly sampled frames. The policy-ranked frames provide high-quality learning signals, while the random frames expand the exploration range and prevent the network from converging to local optima.

4.3 Loss Design

Given the repeat gain $\{\Delta_i\}_{i \in \mathcal{C}}$ computed via AOI for the candidate set $\mathcal{C}$, we train the repeat scoring module to align its predicted scores with these supervision signals through two complementary losses.

Regression Loss. Since the raw scales of the predicted scores and repeat gain may differ significantly across samples, we first standardize both quantities within each sample:

$$\bar{s}_i = \frac{s_i - \mu_s}{\sigma_s + \epsilon}, \quad \bar{\Delta}_i = \frac{\Delta_i - \mu_\Delta}{\sigma_\Delta + \epsilon} \tag{10}$$

where μ_s, σ_s and $\mu_\Delta, \sigma_\Delta$ are the mean and standard deviation of the scores and repeat gain over the candidate set, respectively, and ϵ is a small constant for numerical stability. The regression loss is then defined as:

$$\mathcal{L}_{\text{reg}} = \frac{1}{|\mathcal{C}|} \sum_{i \in \mathcal{C}} \left(\bar{s}_i - \bar{\Delta}_i \right)^2 \tag{11}$$

This formulation encourages the network to produce scores that preserve the relative ordering and magnitude of the repeat gain.

Ranking Loss. While the regression loss aligns scores with repeat gains in a point-wise manner, we additionally introduce a margin-based ranking loss to explicitly enforce the correct ordering between beneficial and detrimental frames. Specifically, we partition the candidate set into positive frames ($\mathcal{C}^+ = \{i \in \mathcal{C} \mid \Delta_i > 0\}$) and negative frames ($\mathcal{C}^- = \{i \in \mathcal{C} \mid \Delta_i < 0\}$), and require that positive frames are scored higher than negative frames by at least a margin m:

$$\mathcal{L}_{\text{rank}} = \frac{1}{|\mathcal{C}^+|} \sum_{i \in \mathcal{C}^+} \frac{1}{|\mathcal{C}^- \cup \mathcal{U}|} \sum_{j \in \mathcal{C}^- \cup \mathcal{U}} \max(0,\ s_j - s_i + m) \tag{12}$$

where $\mathcal{U}$ denotes a randomly sampled subset of non-candidate frames, and m is a margin hyperparameter. By including non-candidate frames as additional negatives, the ranking loss further encourages the network to assign higher scores to frames with positive repeat gain than to the remaining unevaluated frames.

Total Loss. The final training objective is a weighted combination of the two losses:

$$\mathcal{L} = \lambda_{\text{reg}} \cdot \mathcal{L}_{\text{reg}} + \lambda_{\text{rank}} \cdot \mathcal{L}_{\text{rank}} \tag{13}$$

where λ_{reg} and λ_{rank} control the relative importance of the two objectives. The regression loss provides fine-grained score alignment with the repeat gain, while the ranking loss ensures that the critical qualitative distinction is robustly captured.

5 Experiments

5.1 Evaluation Benchmarks

We conduct a comprehensive evaluation of FrameRepeat across five video benchmarks: VideoMME [11], MLVU [45], LongVideoBench [35], LVBench [32], and Video-Holmes [6]. These benchmarks are selected to cover a broad spectrum of video reasoning capabilities. VideoMME comprises videos spanning diverse themes and categorized into short, medium, and long durations, with each video paired with multiple questions to comprehensively assess performance. MLVU consists of seven subtasks designed to evaluate model capabilities from multiple distinct perspectives, including plot understanding, needle-in-a-haystack retrieval, egocentric reasoning, counting, ordering, anomaly recognition, and topic reasoning. LongVideoBench and LVBench focus on long-video understanding, featuring videos ranging up to several hours in length. Video-Holmes serves as a reasoning-intensive benchmark, specifically targeting complex multi-step video reasoning abilities.

5.2 Implementation Details

We train on top of Qwen2.5-VL-7B as the base model. The training data is derived from the multiple-choice subset of LLaVA-Video-178K [43] and the Video-Holmes [6] training set. We perform data filtering by running inference with Qwen2.5-VL-7B five times per sample and discarding those that are answered either entirely correctly or entirely incorrectly across all five runs, resulting in approximately 20K training samples. During training, the video input is uniformly sampled at 128 frames. For the AOI supervision construction, both the top-K and random-K candidate sets use K=16, yielding 32 candidate frames per sample for training efficiency. The repeat scoring module contains only 3.7M trainable parameters, while the base VLM remains entirely frozen throughout training. We set the learning rate to 1×10^{-4} with a per-device batch size of 1. The loss weights are λ_{reg}=1.0 and λ_{rank}=0.1. Training is conducted for one epoch using DeepSpeed ZeRO-2 with BF16 mixed precision.

5.3 Experimental Results

Tab. 1 presents the experimental results of FrameRepeat on five video understanding and reasoning benchmarks. We evaluate on two backbone models, Qwen2.5-VL-7B-Instruct and Qwen3-VL-8B-Thinking, and compare against three baseline frame repeating strategies: CLIP-K, Bottom-K and Random-K.

Across both models, FrameRepeat achieves the best results on all benchmarks. Taking the 128+8 frame setting on Qwen2.5-VL-7B as an example, our method yields improvements of +2.5, +3.7, +2.2, +3.0, and +1.0 on VideoMME, LongVideoBench, LVBench, MLVU, and Video-Holmes, respectively. On Qwen3-VL-8B-Thinking, the improvements are equally significant. In contrast, Bottom-K and Random-K show limited gains and even degrade performance in certain settings, indicating that repeating low-importance or random frames may instead introduce interference. Although the CLIP-K sampling strategy achieves improvements on most benchmarks, it still falls short of our method, demonstrating the effectiveness of our AOI training. The substantial advantage of FrameRepeat over these three strategies demonstrates that our repeat scoring module can accurately predict the importance ranking of each frame, thereby effectively identifying and repeating the frames that truly contribute positively to the model's reasoning. Furthermore, our repeat scoring module is trained only on Qwen2.5-VL-7B, yet achieves consistent improvements on Qwen3-VL-8B-Thinking as well, fully validating the cross-model generalization capability of FrameRepeat.

5.4 Ablation Study

We conduct a series of ablation studies to validate the key design choices in FrameRepeat. Specifically, we examine five aspects: (1) the effect of the repetition count K, (2) the contribution of the ranking loss $\mathcal{L}_{\text{rank}}$, (3) the advantage

Table 1: Main Results. Comparison of five long-video understanding benchmarks. $^{\uparrow x}$ denotes improvement over the corresponding base model at the same frame count and * denotes results obtained under our settings. Best results per model group are in **bold**. We evaluate four frame repeating strategies: CLIP-K (based on CLIP scores), Bottom-K (lowest K scores), Random-K (random repeating), and FrameRepeat (Top-K scores). LVB refers to LongVideoBench.

Model	Frames	VideoMME	LVB	LVBench	MLVU	Video-Holmes
Qwen2.5-VL-72B	768	73.3	60.7	47.3	74.6	–
Qwen3-VL-235B-A22B	768	79.0	–	63.6	83.8	–
InternVL3.5-241B-A28B	–	72.9	67.1	–	78.2	–
Qwen2.5-VL-7B-Instruct						
Base Model*	64	59.6	48.0	32.9	57.1	33.4
+ CLIP-K	64+4	60.2	49.6	33.8	59.4	33.0
+ Bottom-K	64+4	59.6	48.9	32.3	55.9	33.4
+ Random-K	64+4	60.3	48.7	33.2	56.8	33.7
+ FrameRepeat (Ours)	64+4	**61.0**$^{\uparrow 1.4}$	**50.8**$^{\uparrow 2.8}$	**34.3**$^{\uparrow 1.4}$	**60.3**$^{\uparrow 3.2}$	**34.9**$^{\uparrow 1.5}$
Base Model*	128	60.9	51.4	36.0	63.0	35.0
+ CLIP-K	128+8	62.4	53.2	37.5	64.3	35.1
+ Bottom-K	128+8	61.3	52.8	35.4	62.3	35.1
+ Random-K	128+8	61.5	53.0	36.0	61.5	35.0
+ FrameRepeat (Ours)	128+8	**63.4**$^{\uparrow 2.5}$	**55.1**$^{\uparrow 3.7}$	**38.2**$^{\uparrow 2.2}$	**66.0**$^{\uparrow 3.0}$	**36.0**$^{\uparrow 1.0}$
Qwen3-VL-8B-Thinking						
Base Model*	64	62.4	59.2	37.7	62.6	40.0
+ CLIP-K	64+4	62.6	60.3	38.4	63.7	40.2
+ Bottom-K	64+4	62.0	58.5	36.7	63.9	39.3
+ Random-K	64+4	62.3	58.9	38.1	64.5	39.3
+ FrameRepeat (Ours)	64+4	**63.2**$^{\uparrow 0.8}$	**61.5**$^{\uparrow 2.3}$	**39.3**$^{\uparrow 1.6}$	**65.4**$^{\uparrow 2.8}$	**40.5**$^{\uparrow 0.5}$
Base Model*	128	67.0	62.1	40.6	68.2	41.0
+ CLIP-K	128+8	67.0	64.0	42.3	69.4	41.5
+ Bottom-K	128+8	66.3	63.5	41.2	68.3	41.0
+ Random-K	128+8	67.3	63.3	41.6	68.5	40.2
+ FrameRepeat (Ours)	128+8	**67.9**$^{\uparrow 0.9}$	**65.4**$^{\uparrow 3.3}$	**44.3**$^{\uparrow 3.7}$	**71.2**$^{\uparrow 3.0}$	**43.4**$^{\uparrow 2.4}$

of FrameRepeat over frame selection methods, (4) impact on temporal understanding and (5) visualization of repeated frames. Unless otherwise stated, we adopt the same training configuration as the full model.

Effect of Repetition Count. We investigate the impact of varying the number of repeated frames K. As shown in Tab. 4, we compare $K \in \{4, 8, 16\}$ on all five benchmarks. On both models, increasing K from 4 to 8 yields consistent improvements across nearly all benchmarks. However, further increasing to $K =$ 16 shows mixed results: on Qwen2.5-VL, performance drops on all benchmarks compared to $K = 8$; on Qwen3-VL, $K = 16$ surpasses $K = 8$ on MLVU, but declines on the other benchmarks. These results suggest that $K = 8$ provides

Table 2: Effect of Ranking Loss. Performance with and without $\mathcal{L}_{\text{rank}}$ (128+8 frames).

Model	Setting	VideoMME	Video-Holmes
Qwen2.5-VL	w/o $\mathcal{L}_{\text{rank}}$	62.1	35.5
	Full model	**63.4**	**36.0**
Qwen3-VL	w/o $\mathcal{L}_{\text{rank}}$	66.8	42.1
	Full model	**67.9**	**43.4**

Table 3: Comparison with using T-SPO for frame repetition. We use Qwen2.5-VL-7B as the base model and the 128+8 frame setting for tests.

Model	VideoMME	LVB	LVBench
TSPO [24]	62.8	54.3	36.9
FrameRepeat	**63.4**	**55.1**	**38.2**

Table 4: Effect of Repetition Count. Comparison of repeating $K \in \{4, 8, 16\}$ frames on all benchmarks (128+K frames). LVB refers to LongVideoBench.

Model	K	VideoMME	LVB	LVBench	MLVU	Video-Holmes
Qwen2.5-VL	4	62.3	53.9	36.6	64.7	35.2
	8	**63.4**	**55.1**	**38.2**	**66.0**	**36.0**
	16	62.9	53.2	36.0	65.6	34.6
Qwen3-VL	4	67.3	62.8	41.7	70.0	41.4
	8	**67.9**	**65.4**	**44.3**	71.2	**43.4**
	16	67.3	65.0	43.7	**71.3**	43.0

the best overall trade-off, as excessive repetition may introduce redundancy that dilutes the reinforcement effect.

Effect of Ranking Loss. We ablate the ranking loss $\mathcal{L}_{\text{rank}}$ by training with only the regression loss $\mathcal{L}_{\text{reg}}$. As shown in Tab. 2, removing $\mathcal{L}_{\text{rank}}$ leads to consistent performance drops on both models: VideoMME decreases by 1.3 and 1.1 points on Qwen2.5-VL and Qwen3-VL, respectively, and Video-Holmes drops by 0.5 and 1.3 points. This confirms that the ranking loss provides complementary pairwise supervision that helps the network better distinguish frame importance orderings beyond pointwise regression alone.

Comparison with Frame Selection Methods. To further validate our approach, we compare it against a baseline that directly repurposes TSPO [24], a text and event aware frame selection method, for frame repetition. Although frame selection and frame repetition appear related, they serve fundamentally different objectives: the former identifies the most informative frames to **retain**, while the latter seeks to reinforce **temporally critical yet underattended cues**. Simply selecting the highest-scored frames for repetition tends to overemphasize already salient content, rather than compensating for the model's temporal blind spots. Results across multiple benchmarks shown in Tab. 3 consistently demonstrate that our method outperforms the TSPO-based repetition baseline, confirming that a learned repetition strategy is superior to naively reusing scores from a frame selection paradigm.

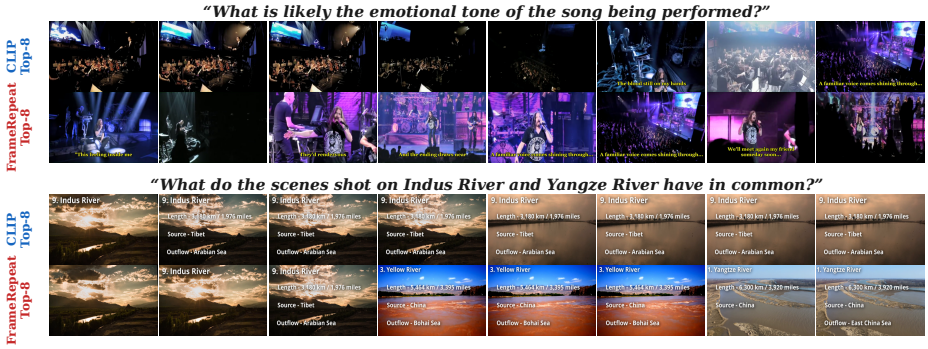

Fig. 4: Comparison of frames selected using FrameRepeat score versus CLIP score.

Table 5: Impact on temporal understanding. Performance on temporal-specific tasks with timestamp-aligned frame repetition (128+8 frames).

Model	VideoMME Temporal		Charades-STA	
	Perception	Reasoning	R@0.5	mIoU
Qwen2.5-VL-7B	69.1	50.3	39.8	37.7
Qwen2.5-VL-7B + FrameRepeat	**71.2**	**57.6**	**42.6**	**40.2**
Qwen3-VL-8B	78.2	55.4	54.2	45.8
Qwen3-VL-8B + FrameRepeat	**81.8**	**59.3**	**55.5**	**48.7**

Impact on Temporal Understanding. Since frame repetition alters the input sequence, a natural concern is whether it disrupts the model's temporal perception. To preserve temporal coherence, we adjust timestamps after repetition: the base 128 frames retain their original timestamps, while each repeated frame is offset from its source frame by 0.01 of the sampling interval, ensuring minimal perturbation to the temporal structure. As shown in Tab. 5, FrameRepeat not only avoids degrading temporal understanding but actually improves it on both temporal perception and reasoning subtasks of VideoMME, as well as on the temporal grounding benchmark Charades-STA [12]. This suggests that reinforcing key frames helps the model anchor temporal events during reasoning.

Visualization. We visualize the frames selected for repetition by FrameRepeat and compare them with those selected using CLIP score. The results are shown in Fig. 4. As can be seen, CLIP score selects frames based solely on global feature similarity, which makes it difficult to capture fine-grained semantic information relevant to the question. In contrast, our frame repeat module, after being trained with AOI, can more accurately assess the importance of each frame for answering the question by incorporating question semantics, thereby selecting frames that truly contain key visual clues. By repeating these frames, the model is better

guided to attend to the important information during its reasoning process, ultimately correcting previously incorrect predictions.

6 Conclusion

In this work, we propose FrameRepeat to mitigate the dilution of visual signals during video reasoning. By repeating key frames at the input stage, we amplify the visual signal strength of these frames, enabling the model to sustain attention to critical information throughout long chain-of-thought reasoning. We design a lightweight repeat scoring module to predict each frame's impact on producing the correct answer, trained via an Add-One-In (AOI) strategy that learns the marginal information contribution of individual frames. Our experimental results demonstrate that FrameRepeat effectively improves model performance on video reasoning tasks and exhibits strong generalization ability, validating the effectiveness of frame repetition in alleviating visual signal dilution during the thinking process.

Acknowledgements

This research was supported by the National Natural Science Foundation of China (Grant No. 62276245).

References

1. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., Ge, W., Guo, Z., Huang, Q., Huang, J., Huang, F., Hui, B., Jiang, S., Li, Z., Li, M., Li, M., Li, K., Lin, Z., Lin, J., Liu, X., Liu, J., Liu, C., Liu, Y., Liu, D., Liu, S., Lu, D., Luo, R., Lv, C., Men, R., Meng, L., Ren, X., Ren, X., Song, S., Sun, Y., Tang, J., Tu, J., Wan, J., Wang, P., Wang, P., Wang, Q., Wang, Y., Xie, T., Xu, Y., Xu, H., Xu, J., Yang, Z., Yang, M., Yang, J., Yang, A., Yu, B., Zhang, F., Zhang, H., Zhang, X., Zheng, B., Zhong, H., Zhou, J., Zhou, F., Zhou, J., Zhu, Y., Zhu, K.: Qwen3-vl technical report (2025), https://arxiv.org/abs/2511.21631
2. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., Zhong, H., Zhu, Y., Yang, M., Li, Z., Wan, J., Wang, P., Ding, W., Fu, Z., Xu, Y., Ye, J., Zhang, X., Xie, T., Cheng, Z., Zhang, H., Yang, Z., Xu, H., Lin, J.: Qwen2.5-vl technical report (2025), https://arxiv.org/abs/2502.13923
3. Cao, M., Zhao, H., Zhang, C., Chang, X., Reid, I., Liang, X.: Ground-r1: Incentivizing grounded visual reasoning via reinforcement learning. arXiv preprint arXiv:2505.20272 (2025)
4. Chen, Y., Li, L., Xi, T., Zeng, L., Wang, J.: Perception before reasoning: Two-stage reinforcement learning for visual reasoning in vision-language models. arXiv preprint arXiv:2509.13031 (2025)
5. Chen, Z., Tao, J., Li, R., Hu, Y., Chen, R., Yang, Z., Yu, X., Jing, H., Zhang, M., Shao, S., et al.: Omnivideo-r1: Reinforcing audio-visual reasoning with query intention and modality attention. arXiv preprint arXiv:2602.05847 (2026)

6. Cheng, J., Ge, Y., Wang, T., Ge, Y., Liao, J., Shan, Y.: Video-holmes: Can mllm think like holmes for complex video reasoning? arXiv preprint arXiv:2505.21374 (2025)
7. Ding, Y., Zhang, Y., Lai, X., Chu, R., Yang, Y.: Videozoomer: Reinforcement-learned temporal focusing for long video reasoning. arXiv preprint arXiv:2512.22315 (2025)
8. Dong, Y., Liu, Z., Sun, H.L., Yang, J., Hu, W., Rao, Y., Liu, Z.: Insight-v: Exploring long-chain visual reasoning with multimodal large language models. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 9062–9072 (2025)
9. Fei, H., Wu, S., Ji, W., Zhang, H., Zhang, M., Lee, M.L., Hsu, W.: Video-of-thought: Step-by-step video reasoning from perception to cognition. arXiv preprint arXiv:2501.03230 (2024)
10. Feng, K., Gong, K., Li, B., Guo, Z., Wang, Y., Peng, T., Wu, J., Zhang, X., Wang, B., Yue, X.: Video-r1: Reinforcing video reasoning in mllms. arXiv preprint arXiv:2503.21776 (2025)
11. Fu, C., Dai, Y., Luo, Y., Li, L., Ren, S., Zhang, R., Wang, Z., Zhou, C., Shen, Y., Zhang, M., et al.: Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 24108–24118 (2025)
12. Gao, J., Sun, C., Yang, Z., Nevatia, R.: Tall: Temporal activity localization via language query. In: Proceedings of the IEEE international conference on computer vision. pp. 5267–5275 (2017)
13. He, Z., Qu, X., Li, Y., Huang, S., Liu, D., Cheng, Y.: Framethinker: Learning to think with long videos via multi-turn frame spotlighting. arXiv preprint arXiv:2509.24304 (2025)
14. Hong, J., Zhao, C., Zhu, C., Lu, W., Xu, G., Yu, X.: Deepeyesv2: Toward agentic multimodal model. arXiv preprint arXiv:2511.05271 (2025)
15. Jian, P., Wu, J., Sun, W., Wang, C., Ren, S., Zhang, J.: Look again, think slowly: Enhancing visual reflection in vision-language models. In: Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing. pp. 9262–9281 (2025)
16. Jiang, J., Ma, C., Song, X., Zhang, H., Luo, J.: Corvid: Improving multimodal large language models towards chain-of-thought reasoning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3034–3046 (2025)
17. Li, J., Yin, H., Tan, W., Chen, J., Xu, B., Qu, Y., Chen, Y., Ju, J., Luo, Z., Luan, J.: Revisor: Beyond textual reflection, towards multimodal introspective reasoning in long-form video understanding. arXiv preprint arXiv:2511.13026 (2025)
18. Li, X., Yan, Z., Meng, D., Dong, L., Zeng, X., He, Y., Wang, Y., Qiao, Y., Wang, Y., Wang, L.: Videochat-r1: Enhancing spatio-temporal perception via reinforcement fine-tuning. arXiv preprint arXiv:2504.06958 (2025)
19. Liu, S., Zhuge, M., Zhao, C., Chen, J., Wu, L., Liu, Z., Zhu, C., Cai, Z., Zhou, C., Liu, H., et al.: Videoauto-r1: Video auto reasoning via thinking once, answering twice. arXiv preprint arXiv:2601.05175 (2026)
20. Liu, Z., Sun, Z., Zang, Y., Dong, X., Cao, Y., Duan, H., Lin, D., Wang, J.: Visual-rft: Visual reinforcement fine-tuning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2034–2044 (2025)
21. Luo, M., Xue, Z., Dimakis, A., Grauman, K.: When thinking drifts: Evidential grounding for robust video reasoning. arXiv preprint arXiv:2510.06077 (2025)

22. Ouyang, K., Liu, Y., Wu, H., Liu, Y., Zhou, H., Zhou, J., Meng, F., Sun, X.: Spacer: Reinforcing mllms in video spatial reasoning. arXiv preprint arXiv:2504.01805 (2025)
23. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
24. Tang, C., Han, Z., Sun, H., Zhou, S., Zhang, X., Wei, X., Yuan, Y., Zhang, H., Xu, J., Sun, H.: Tspo: Temporal sampling policy optimization for long-form video language understanding. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 40, pp. 9368–9376 (2026)
25. Tao, Z., Wang, S., Hua, Y., Cao, H., Xu, L.: Dig: Differential grounding for enhancing fine-grained perception in multimodal large language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1695–1705 (2026)
26. Tao, Z., Zhang, F., Ding, Z., Wang, S., Sun, X., Hua, Y., Cao, H., Xu, L.: Locus: Local visual cue search for enhancing fine-grained perception in multimodal large language models. arXiv preprint arXiv:2606.16586 (2026)
27. Team, K., Du, A., Yin, B., Xing, B., Qu, B., Wang, B., Chen, C., Zhang, C., Du, C., Wei, C., et al.: Kimi-vl technical report. arXiv preprint arXiv:2504.07491 (2025)
28. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
29. Wang, H., Liu, H., Liu, X., Du, C., Kawaguchi, K., Wang, Y., Pang, T.: Fostering video reasoning via next-event prediction. arXiv preprint arXiv:2505.22457 (2025)
30. Wang, Q., Yu, Y., Yuan, Y., Mao, R., Zhou, T.: Videorft: Incentivizing video reasoning capability in mllms via reinforced fine-tuning. arXiv preprint arXiv:2505.12434 (2025)
31. Wang, S., Hua, Y., Tao, Z., Cao, H., Xu, L.: Dynamic token compression for efficient video understanding through reinforcement learning. arXiv preprint arXiv:2603.26365 (2026)
32. Wang, W., He, Z., Hong, W., Cheng, Y., Zhang, X., Qi, J., Ding, M., Gu, X., Huang, S., Xu, B., et al.: Lvbench: An extreme long video understanding benchmark. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 22958–22967 (2025)
33. Wang, W., Gao, Z., Gu, L., Pu, H., Cui, L., Wei, X., Liu, Z., Jing, L., Ye, S., Shao, J., Wang, Z., Chen, Z., Zhang, H., Yang, G., Wang, H., Wei, Q., Yin, J., Li, W., Cui, E., Chen, G., Ding, Z., Tian, C., Wu, Z., Xie, J., Li, Z., Yang, B., Duan, Y., Wang, X., Hou, Z., Hao, H., Zhang, T., Li, S., Zhao, X., Duan, H., Deng, N., Fu, B., He, Y., Wang, Y., He, C., Shi, B., He, J., Xiong, Y., Lv, H., Wu, L., Shao, W., Zhang, K., Deng, H., Qi, B., Ge, J., Guo, Q., Zhang, W., Zhang, S., Cao, M., Lin, J., Tang, K., Gao, J., Huang, H., Gu, Y., Lyu, C., Tang, H., Wang, R., Lv, H., Ouyang, W., Wang, L., Dou, M., Zhu, X., Lu, T., Lin, D., Dai, J., Su, W., Zhou, B., Chen, K., Qiao, Y., Wang, W., Luo, G.: Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency (2025), https://arxiv.org/abs/2508.18265
34. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., Zhou, D.: Chain-of-thought prompting elicits reasoning in large language models (2023), https://arxiv.org/abs/2201.11903

35. Wu, H., Li, D., Chen, B., Li, J.: Longvideobench: A benchmark for long-context interleaved video-language understanding. Advances in Neural Information Processing Systems **37**, 28828–28857 (2024)
36. Wu, P., Zhang, Y., Diao, H., Li, B., Lu, L., Liu, Z.: Visual jigsaw post-training improves mllms. arXiv preprint arXiv:2509.25190 (2025)
37. Xiao, W., Gan, L., Dai, W., He, W., Huang, Z., Li, H., Shu, F., Yu, Z., Zhang, P., Jiang, H., et al.: Fast-slow thinking for large vision-language model reasoning. arXiv e-prints pp. arXiv–2504 (2025)
38. Xing, L., Dong, X., Zang, Y., Cao, Y., Liang, J., Huang, Q., Wang, J., Wu, F., Lin, D.: Caprl: Stimulating dense image caption capabilities via reinforcement learning. arXiv preprint arXiv:2509.22647 (2025)
39. Xu, G., Jin, P., Wu, Z., Li, H., Song, Y., Sun, L., Yuan, L.: Llava-cot: Let vision language models reason step-by-step. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2087–2098 (2025)
40. Yang, S., Niu, Y., Liu, Y., Ye, Y., Lin, B., Yuan, L.: Look-back: Implicit visual re-focusing in mllm reasoning. arXiv preprint arXiv:2507.03019 (2025)
41. Yang, Y., He, X., Pan, H., Jiang, X., Deng, Y., Yang, X., Lu, H., Yin, D., Rao, F., Zhu, M., et al.: R1-onevision: Advancing generalized multimodal reasoning through cross-modal formalization. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2376–2385 (2025)
42. Yao, H., Huang, J., Wu, W., Zhang, J., Wang, Y., Liu, S., Wang, Y., Song, Y., Feng, H., Shen, L., et al.: Mulberry: Empowering mllm with o1-like reasoning and reflection via collective monte carlo tree search. arXiv preprint arXiv:2412.18319 (2024)
43. Zhang, Y., Wu, J., Li, W., Li, B., Ma, Z., Liu, Z., Li, C.: Llava-video: Video instruction tuning with synthetic data (2025), https://arxiv.org/abs/2410.02713
44. Zheng, Z., Yang, M., Hong, J., Zhao, C., Xu, G., Yang, L., Shen, C., Yu, X.: Deepeyes: Incentivizing" thinking with images" via reinforcement learning. arXiv preprint arXiv:2505.14362 (2025)
45. Zhou, J., Shu, Y., Zhao, B., Wu, B., Xiao, S., Yang, X., Xiong, Y., Zhang, B., Huang, T., Liu, Z.: Mlvu: A comprehensive benchmark for multi-task long video understanding. arXiv e-prints pp. arXiv–2406 (2024)