

# VideoSSR: Video Self-Supervised Reinforcement Learning

Zefeng He<sup>1,2</sup> Xiaoye Qu<sup>1\*</sup> Yafu Li<sup>3</sup> Siyuan Huang<sup>4</sup> Daizong Liu<sup>5</sup> Yu Cheng<sup>3\*</sup>

<sup>1</sup>Shanghai Artificial Intelligence Laboratory, <sup>2</sup>Nanjing University

<sup>3</sup>The Chinese University of Hong Kong

<sup>4</sup>Shanghai Jiao Tong University, <sup>5</sup>Wuhan University

## Abstract

Reinforcement Learning with Verifiable Reward (RLVR) has substantially advanced the video understanding capabilities of Multimodal Large Language Models (MLLMs). However, the rapid progress of MLLMs is outpacing the complexity of existing video datasets, while the manual annotation of new, high-quality data remains prohibitively expensive. This work investigates a pivotal question: Can the rich, intrinsic information within videos be harnessed to self-generate high-quality, verifiable training data? To investigate this problem, we first introduce three self-supervised pretext tasks for video understanding: Anomaly Grounding, Object Counting, and Temporal Jigsaw. To validate the difficulty of these tasks, we construct the Video Intrinsic Understanding Benchmark (VIUBench), revealing that current state-of-the-art MLLMs struggle significantly on these tasks. Building upon these pretext tasks, we develop the VideoSSR-30K dataset and propose VideoSSR, a novel video self-supervised reinforcement learning framework for RLVR. Extensive experiments across 17 benchmarks, spanning four major video domains (General Video QA, Long Video QA, Temporal Grounding, and Complex Reasoning), demonstrate that our VideoSSR consistently enhances model performance, yielding an average improvement of over 5%. These results establish VideoSSR as a potent foundational framework for developing more advanced video understanding in MLLMs. The code is available at <https://github.com/lcqysl/VideoSSR>.

## 1. Introduction

In past years, Multimodal Large Language Models (MLLMs) have achieved remarkable progress in the field of video understanding [2, 3, 9, 34, 35, 39, 45, 49]. Benefiting from recent Reinforcement Learning with Verifiable Reward (RLVR) [11, 16, 26, 43, 44, 64, 66], the performance of MLLMs has been further improved. A corner-

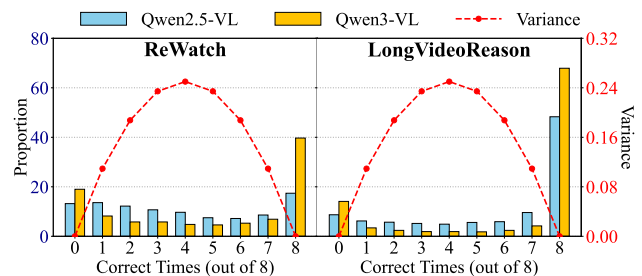


Figure 1. **Distribution of answer correctness on ReWatch and LongVideoReason.** Across both models and datasets, a vast majority of questions yield a bimodal outcome, resulting in either zero or eight correct answers. This zero variance issue is notably more pronounced for the more powerful Qwen3-VL model.

stone of the RLVR approach is the availability of video datasets with verifiable answers. To obtain the verifiable answers, existing datasets, such as LongVideoReason [8] and ReWatch [64], utilize multi-agent collaboration to construct high-quality datasets with verifiable answers.

Although current datasets have effectively enhanced the performance of models like Qwen2.5-VL [3], significant limitations arise when applying them to more powerful models, such as the recent Qwen3-VL [39]. First, for highly capable models, many questions in existing datasets lack sufficient complexity. To illustrate this, we generate eight independent responses per question using Qwen2.5-VL [3] and Qwen3-VL [39]. As shown in Figure 1, a vast majority of questions yield a perfect score where all eight responses are correct, indicating they are insufficiently challenging. Second, the previous multi-agent annotation process introduces systemic biases and artifacts, which create flawed or spurious reward signals for RLVR, particularly when the annotator models are less capable than the target models. This is evidenced by another large portion of questions where all generated responses are incorrect, suggesting either intractable difficulty or biased ground truths. The resulting bimodal distribution of scores, with most questions exhibiting zero variance, offers an ineffective learning signal for GRPO [16, 43] training in RLVR. Consequently, training

\*Corresponding authors

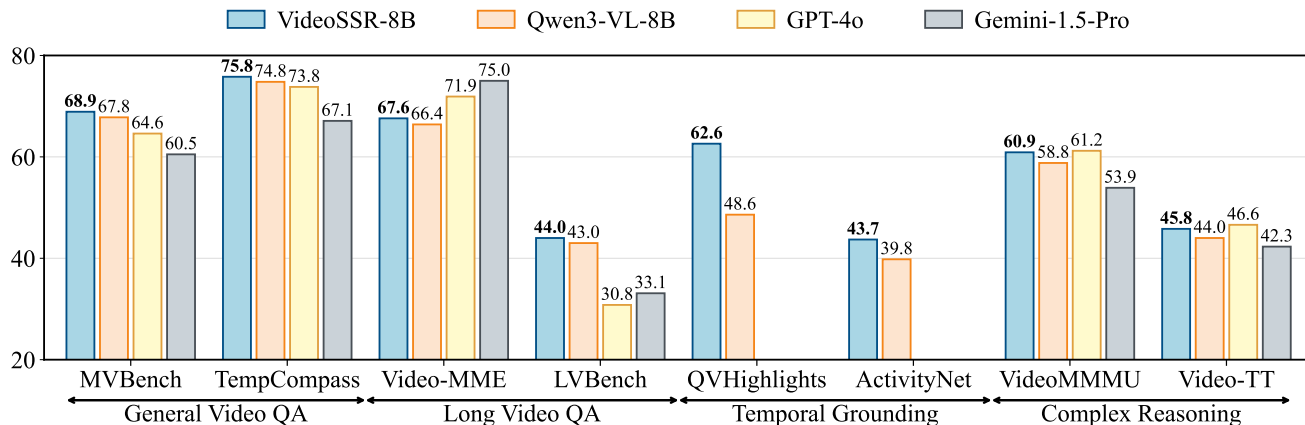


Figure 2. **Performance comparison on four video tasks.** VideoSSR-8B consistently outperforms Qwen3-VL-8B across four video tasks.

advanced models on such data yields marginal gains or even performance degradation (Section 4.2).

Compounding the above issues is the prohibitive cost of manual annotation for video. This predicament, however, points to a compelling alternative: can the rich, intrinsic information within videos be harnessed to construct high-quality, verifiable questions for RLVR? Inspired by traditional video self-supervised learning [12, 32, 33, 57], we first design three self-supervised pretext tasks with parametrically scalable difficulty, including Anomaly Grounding, Object Counting, and Temporal Jigsaw, to generate verifiable questions for video understanding. To validate the difficulty of these tasks, we construct **Video Intrinsic Understanding Bench (VIUBench)** and observe that questions targeting the intrinsic properties of the video itself remain profoundly challenging, even for leading closed-source models like GPT-5 [35].

Building on this insight, we introduce **VideoSSR**, a new **Video Self-Supervised Reinforcement** learning framework to enhance the video understanding of MLLM. We construct the VideoSSR-30K dataset using the aforementioned pretext tasks, which is entirely independent of human or MLLM annotations. This dataset is subsequently utilized to train our VideoSSR-8B model with GRPO. Moreover, to overcome the challenge of sparse reward signals arising from the inherent difficulty of these tasks, we design corresponding smooth reward functions for each pretext task to ensure efficient and stable RLVR training.

To validate the generalization capability of VideoSSR, we conduct extensive experiments on 17 benchmarks spanning four main video tasks: General Video QA, Long Video QA, Temporal Grounding, and Complex Reasoning. The results show that our proposed VideoSSR achieves consistent performance improvements across all benchmarks and under three different input frame settings, demonstrating an average gain of over 5% with 48 input frames.

In summary, our main contributions are fourfold:

- We generate verifiable training data for RLVR that harnesses intrinsic video signals. This self-supervised paradigm circumvents the prohibitive costs and inherent biases of prevailing multi-agent and manual annotation, thereby addressing a critical bottleneck in scaling MLLMs for video understanding.
- We introduce three self-supervised pretext tasks with parametrically scalable difficulty and construct VIUBench, revealing profound limitations in state-of-the-art MLLMs for intrinsic video understanding.
- We introduce VideoSSR, a self-supervised reinforcement learning framework for video RLVR training and construct VideoSSR-30K dataset. To facilitate efficient and stable RLVR training in VideoSSR, we further design three tailored smooth reward functions.
- Extensive experiments across 17 benchmarks demonstrate the superior generalization capability of VideoSSR. Our method consistently achieves significant performance improvements, establishing it as a foundational framework for advancing video understanding.

## 2. Related Works

### 2.1. Reinforcement Learning for MLLMs

Reinforcement Learning with Verifiable Reward (RLVR) [16, 43] has been shown to significantly enhance the capabilities of language models, a success that has been rapidly extended to MLLMs [7, 8, 10, 18, 26, 38, 59]. For instance, Video-R1 [11] leverages existing Video QA datasets [37, 53, 56, 61, 67] to bolster performance on Video QA tasks. Time-R1 [50] utilizes datasets with precise timestamp annotations to improve Temporal Grounding capability. SpaceR [36] automatically generates verifiable questions from the geometric and semantic ground truths of 3D scenes [4, 30, 60], enhancing the model’s spatial

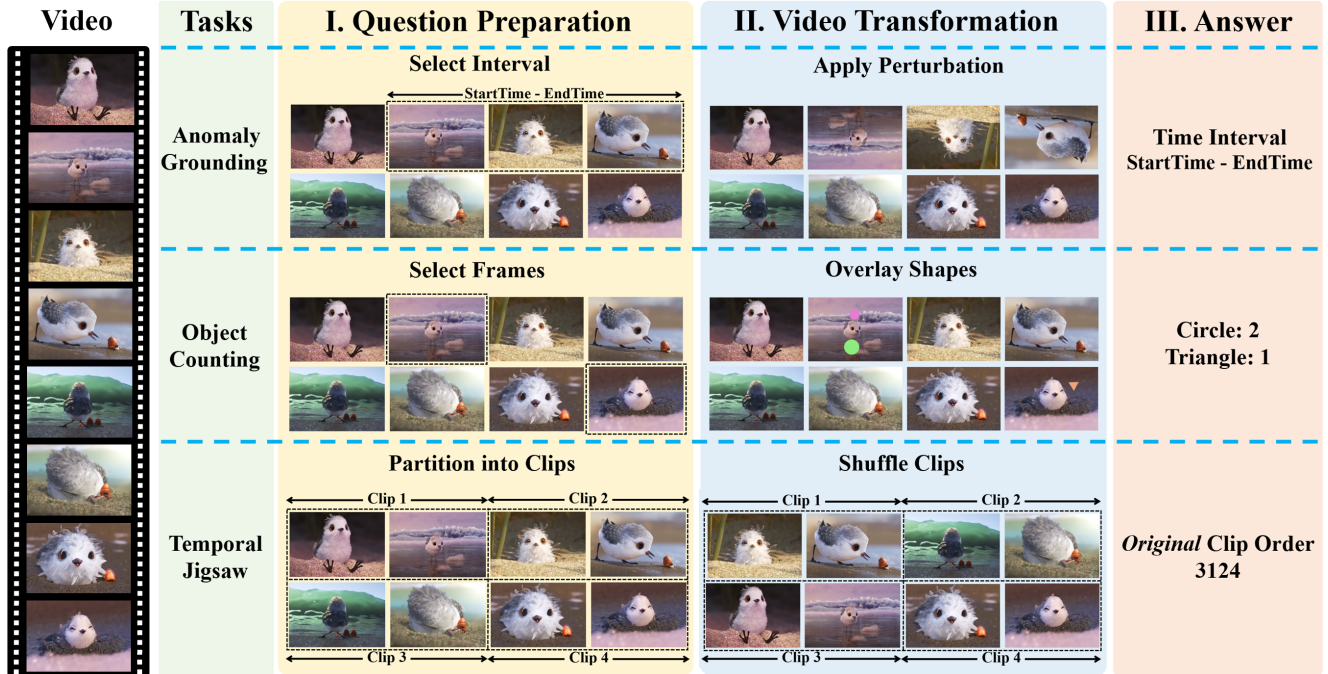


Figure 3. **An overview of our three self-supervised pretext tasks.** (a) **Anomaly Grounding:** A temporal segment is perturbed (e.g., via rotation), and the task is to identify the start and end timestamps of this anomaly. (b) **Object Counting:** Procedurally generated shapes are overlaid onto selected frames, and the task is to count the total number of each shape type. (c) **Temporal Jigsaw:** The video is divided into clips which are then shuffled. The task is to predict the original temporal order of the segments.

reasoning abilities. ReWatch-R1 [64] leverages multi-agent collaboration to construct high-quality reasoning datasets, thereby advancing its capabilities in complex reasoning. Despite these diverse data sourcing strategies, several fundamental limitations persist. The reliance on external annotations often introduces significant bias. Meanwhile, many approaches often specialize in enhancing a single capability, which can limit their broader generalization.

## 2.2. Self-supervised learning for Video

Self-supervised learning [12, 23, 27, 28, 32, 33, 42, 47, 57] for video aims to learn effective spatio-temporal representations from unlabeled video data. A dominant approach involves designing pretext tasks that capitalize on the inherent properties of video. For instance, early works leverage tasks such as video jigsaw puzzles [1, 22, 32, 46, 57] to learn representations. Similarly, recent research has employed the jigsaw puzzle task [19, 51, 55, 63] to facilitate the reinforcement learning of MLLMs. However, this reliance on a single task means the full potential of the self-supervised paradigm for cultivating comprehensive video understanding remains underexplored. To instill stronger generalization capabilities in MLLMs, our work expands beyond a singular pretext task by leveraging a diverse collection of self-supervised challenges.

## 3. Method

Considering there is rich information in the video, in this paper, we explore leveraging the intrinsic information within the video itself to construct high-quality questions with scalable difficulty. To investigate it, we begin by designing three pretext tasks.

### 3.1. Pretext Tasks

In this section, we introduce three pretext tasks, including Anomaly Grounding, Object Counting, and Temporal Jigsaw. These tasks share a common design philosophy, namely, they can generate verifiable question-answer pairs directly from raw videos, independent of any human or model-generated annotations. Furthermore, the difficulty of these pairs can be parametrically controlled. The overall process for these three tasks is illustrated in Figure 3.

#### 3.1.1. Anomaly Grounding

This task assesses the model’s ability to localize temporal segments that violate natural video dynamics. Let a video be represented as a sequence of frames  $V = \{f_1, f_2, \dots, f_T\}$ , with a total duration of  $D$  seconds. We first randomly select a temporal interval  $[t_s, t_e] \subseteq [0, D]$ , where  $t_s$  and  $t_e$  are the start and end timestamps, respectively. This interval corresponds to a contiguous segment of

frames  $S = \{f_i \mid \text{timestamp}(f_i) \in [t_s, t_e]\}$ .

Next, we apply a perturbation function  $\mathcal{P}$  to this segment to create a perturbed version,  $S' = \mathcal{P}(S)$ . The function  $\mathcal{P}$  is sampled from a set of predefined transformations targeting different core capabilities:

- **Fine-grained Perception:** e.g., swapping the red and blue color channels for every frame in  $S$ .
- **Spatial Perception:** e.g., rotating every frame in  $S$  by 180 degrees.
- **Temporal Perception:** e.g., randomly shuffling the frame order within  $S$ .

The final video  $V'$  is constructed by replacing the original segment  $S$  with its perturbed counterpart  $S'$ . The model is then provided with the modified video  $V'$  and is tasked to identify the anomalous interval by predicting its start and end timestamps,  $(t_s, t_e)$ .

### 3.1.2. Object Counting

This task targets the model’s fine-grained perception and counting abilities. We define a set of primitive geometric shapes  $\mathcal{C} = \{c_1, c_2, \dots, c_K\}$ , such as circles, rectangles, and triangles, which can be procedurally generated. For a given video  $V$ , we randomly select a subset of frames  $\mathcal{F}_{sub} \subset V$ . For each frame  $f_i \in \mathcal{F}_{sub}$ , we synthesize a set of objects  $O_i$ , where each object  $o \in O_i$  is an instance of a shape class from  $\mathcal{C}$  with randomized attributes (size, color, rotation, position). These modified frames, denoted as  $f'_i$ , are then used to create the final video  $V'$  by replacing their original counterparts. The ground truth is a vector of counts  $\mathbf{n} = [N_1, N_2, \dots, N_K]$ , where each element  $N_k$  is the total number of occurrences of shape  $c_k$ :

$$N_k = \sum_{f_i \in \mathcal{F}_{sub}} |\{o \in O_i \mid \text{type}(o) = c_k\}| \quad (1)$$

Given  $V'$ , the model is required to output the counts for each shape category.

### 3.1.3. Temporal Jigsaw

This task is designed to evaluate the model’s temporal perception, specifically its understanding of temporal coherence and event ordering. We partition the video  $V$  into  $n$  contiguous, non-overlapping segments of equal duration,  $V = [S_1, S_2, \dots, S_n]$ . We then generate a random permutation  $\pi$  of the indices  $\{1, 2, \dots, n\}$ . A new video  $V'$  is created by reordering the segments according to this permutation:

$$V' = [S_{\pi(1)}, S_{\pi(2)}, \dots, S_{\pi(n)}] \quad (2)$$

The model is presented with the shuffled video  $V'$  and is tasked with restoring the original temporal order. To achieve this, it must predict a sequence of indices that correctly reorders the shuffled segments. This target sequence

is the inverse of the permutation  $\pi$  that was used for shuffling. The answer is therefore the sequence defined by  $\pi^{-1}$ :

$$\text{Answer} = (\pi^{-1}(1), \pi^{-1}(2), \dots, \pi^{-1}(n)) \quad (3)$$

## 3.2. Video Intrinsic Understanding Benchmark

After defining the above three tasks, a critical question arises: are these pretext tasks sufficiently challenging for state-of-the-art MLLMs? To investigate this, we construct the **Video Intrinsic Understanding Bench (VIUBench)**, which systematically evaluates a model’s ability to comprehend intrinsic video properties across three core axes: Fine-grained Perception, Spatial Perception, and Temporal Perception. The benchmark is composed of 2700 video question-answer pairs generated from our three pretext tasks. The proportional distribution of data across these tasks is illustrated in the left panel of Figure 4.

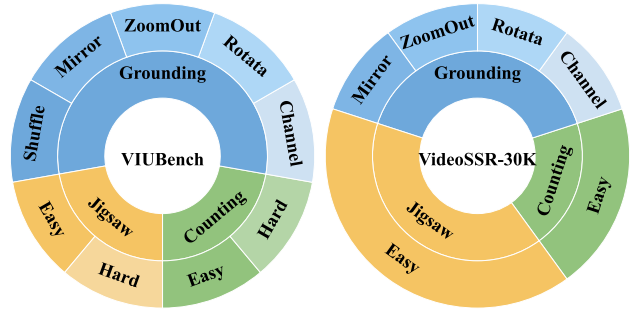


Figure 4. **Task distribution in VIUBench and VideoSSR-30K.** The left panel illustrates the proportional data distribution across our three pretext tasks and their subtypes for VIUBench. The right panel shows the corresponding composition of VideoSSR-30K.

**Anomaly Grounding.** For the Anomaly Grounding task, we select five representative perturbation types from a larger pool of 14 (detailed in Supplementary Material 2.1). The selected types include: (1) swapping the red and blue color channels, (2) rotation by 180 degrees, (3) zooming out, (4) horizontal mirroring, and (5) shuffling the intra-segment frame order. We compute the Mean Intersection over Union (mIoU) between the predicted and ground-truth temporal intervals as the performance score.

**Object Counting.** For the Object Counting task, we use three primitive shapes (circles, rectangles, and triangles) and configure two difficulty levels:

- **Easy:** Objects are overlaid onto a maximum of three frames, with no more than three instances of any single shape type appearing in any given frame.
- **Hard:** The constraints are increased to a maximum of four frames and up to four instances per shape per frame.

A score of 1 is awarded for a specific shape type if the predicted count is exactly equal to the ground-truth count, and

Table 1. **Performance comparison on VIUBench.** The benchmark assesses three core abilities (Fine-Grained Perception, Spatial Perception, and Temporal Perception) via our three pretext tasks (Object Counting, Anomaly Grounding, and Temporal Jigsaw). For both open source and closed source models, the top result is shown in bold, and the second-best is underlined.

Ability	Fine-Grained Perception			Spatial Perception			Temporal Perception			Average			
Task	Object Counting		Anomaly Grounding			Temporal Jigsaw			Counting	Grounding	Jigsaw	Overall	
Type	Easy	Hard	Channel	Rotate	ZoomOut	Mirror	Shuffle	Easy	Hard	-	-	-	-
<i>Random Guess</i>													
Random Guess	11.1	6.3	25.9	25.4	25.4	25.2	25.2	0.1	0.0	8.7	25.4	0.1	16.1
<i>Closed Source Models</i>													
GPT-5 [35]	<b>88.4</b>	<b>70.3</b>	<u>82.6</u>	<u>81.8</u>	<b>56.5</b>	<u>48.9</u>	<u>34.1</u>	<b>39.0</b>	<b>27.0</b>	<b>79.4</b>	<u>60.8</u>	<b>33.0</b>	<b>58.7</b>
Gemini-2.5-Pro [9]	80.8	61.3	<b>84.6</b>	<b>82.3</b>	<u>51.0</u>	<b>55.7</b>	<b>52.1</b>	25.3	17.7	<u>71.1</u>	<b>65.1</b>	21.5	56.7
Gemini-2.5-Flash [9]	35.7	22.4	75.8	73.7	28.5	30.2	28.6	8.3	4.0	29.1	47.4	6.2	34.1
Seed1.5-VL [15]	72.3	52.0	79.0	70.7	19.4	31.4	24.1	20.7	9.3	62.6	44.9	15.0	42.2
<i>Open Source Models</i>													
Qwen2.5-VL-7B-Instruct [3]	11.3	5.3	6.4	13.7	8.1	5.2	7.0	0.7	0.0	8.3	8.1	0.3	6.4
VideoJigsaw-7B [55]	12.1	5.4	1.5	4.5	1.2	1.1	2.0	<u>20.3</u>	<u>5.0</u>	8.8	2.1	<u>12.7</u>	5.9
Qwen3-VL-8B-Instruct [39]	13.8	7.7	50.1	53.4	21.4	13.6	14.1	1.3	0.0	10.7	30.5	0.7	19.5
Qwen3-VL-32B-Instruct [39]	20.1	13.0	66.1	63.0	17.6	29.4	18.9	1.3	0.0	16.6	39.0	0.7	25.5
Qwen3-VL-235B-A22B-Instruct [39]	23.8	14.8	<u>68.9</u>	<u>67.8</u>	<u>32.9</u>	28.3	<u>26.8</u>	7.7	3.3	19.3	<u>45.0</u>	5.5	30.5
GLM-4.5V [20]	<b>59.1</b>	<b>45.4</b>	66.4	61.0	21.2	<u>29.8</u>	15.3	11.0	3.3	<b>52.3</b>	38.7	7.2	<u>34.7</u>
InternVL-3.5-8B [49]	15.2	9.6	25.9	42.4	6.1	9.8	3.1	0.0	0.0	12.4	17.5	0.0	12.5
InternVL-3.5-38B [49]	28.0	15.9	47.0	54.5	9.6	18.2	12.4	0.0	0.0	21.9	28.3	0.0	20.6
<b>VideoSSR-8B (Ours)</b>	<u>29.0</u>	<u>24.6</u>	<b>88.7</b>	<b>89.0</b>	<b>94.4</b>	<b>67.8</b>	<b>41.0</b>	<b>24.3</b>	<b>8.0</b>	<u>26.8</u>	<b>76.2</b>	<b>16.2</b>	<b>51.9</b>

0 otherwise. The final task score is the average of these binary scores across all shape types.

**Temporal Jigsaw.** The Temporal Jigsaw task is configured with two difficulty settings based on the number of segments the video is partitioned into:

- **Easy:** The video is partitioned into 6 segments.
- **Hard:** The video is partitioned into 8 segments.

A score is awarded only if the predicted sequence of segments is identical to the ground-truth permutation.

As shown in Table 1, we evaluate a suite of powerful MLLMs on VIUBench. Our findings reveal that this benchmark poses a significant challenge even for the most advanced models. Notably, even a strong closed-source model like GPT-5 [35] only achieves a modest average score of 58.7. The performance of open-source models is even more limited. For instance, Qwen3-VL-8B attains an average score of just 19.5. These results underscore a critical insight that understanding and reasoning about intrinsic video properties, such as fine-grained details and temporal coherence, remains a substantial bottleneck for current MLLMs. This highlights the effectiveness of VIUBench in exposing the limitations of existing models and validates its role as a challenging benchmark for future research.

More importantly, our experiments with VIUBench reveal a key advantage of these pretext tasks: the difficulty of the generated questions can be easily scaled by adjusting simple parameters. For instance, in the Object Counting task, switching from the “Easy” to the “Hard” configuration caused the score of GPT-5 to drop sharply from 88.4 to 70.3. A similar trend is observed in the Temporal Jigsaw task. By increasing the number of video segments from six to eight,

the model’s score plummeted from 39.0 to 27.0. Even for Video Jigsaw [55], a model specifically trained on jigsaw tasks, its performance decreases significantly from 20.3 to 5.0 under the same conditions. To sum up, all these findings demonstrate that our method can dynamically generate tasks that challenge powerful MLLMs. The ability to parametrically control task difficulty ensures that VIUBench can remain a relevant and challenging benchmark for evaluating the continuous advancements of future models.

### 3.3. Video Self-Supervised Reinforcement Learning

Motivated by the insights from our proposed VIUBench, we introduce a novel framework that leverages **Video Self-Supervised Reinforcement learning (VideoSSR)** to enhance the generalization of MLLMs. To perform RLVR, we first construct the **VideoSSR-30K** dataset, which consists of the aforementioned three pretext tasks. The proportional distribution of VideoSSR-30K dataset is detailed in the right panel of Figure 4. A detailed breakdown of the data composition is provided in Supplementary Material 1.3.

For RLVR training, we employ GRPO [16, 43]. We do not use recent variants of GRPO [62, 69], as our primary focus is on the data itself. Our reward function is based solely on answer correctness. However, the inherent difficulty of these tasks poses a problem for RLVR training: using a strict reward function often results in sparse rewards, leading to inefficient and unstable training.

To address this challenge, we design a specific smooth reward function for each task to provide a denser and more informative learning signal. The design of these functions is primarily based on the concept of relative error rate.

Table 2. Performance comparison on General Video QA and Long Video QA tasks.

Model	Frames	General Video QA				Long Video QA			
		MVBench	TempCompass	AoTBench	VinoGround	Video-MME	LVBench	LongVideoBench	CGBench
<i>Closed Source Models</i>									
GPT-4o [34]	-	64.6	73.8	63.4	54	71.9	30.8	62.0	45.2
Gemini-1.5-Pro [45]	-	60.5	67.1	58.3	35.8	75.0	33.1	58.6	37.2
<i>Open Source Models</i>									
Qwen3-VL-8B-Instruct [39]	32	66.9	74.9	59.8	45.2	64.1	39.6	58.6	40.1
	48	67.6	74.8	60.2	45.0	66.0	41.5	60.0	41.7
	64	67.8	74.8	60.7	45.0	66.4	43.0	61.3	42.6
<b>VideoSSR-8B (Ours)</b>	32	68.6(+1.7)	75.7(+0.8)	60.5(+0.7)	55.6(+10.4)	65.2(+1.1)	41.5(+1.9)	59.6(+1.0)	40.2(+0.1)
	48	68.8(+1.2)	75.8(+1.0)	61.7(+1.5)	55.6(+10.6)	66.7(+0.7)	42.9(+1.4)	61.1(+1.1)	42.5(+0.8)
	64	68.9(+1.1)	75.7(+0.9)	61.8(+1.1)	55.6(+10.6)	67.6(+1.2)	44.0(+1.0)	61.5(+0.2)	43.4(+0.8)

**Anomaly Grounding.** For the temporal grounding of anomalies, the Intersection over Union (mIoU) naturally serves as a smooth reward signal. It provides a score between 0 and 1 that reflects the degree of overlap between the predicted and ground-truth temporal segments. Let  $T_{\text{pred}}$  and  $T_{\text{gt}}$  be the predicted and ground-truth intervals, respectively. The reward  $R_{\text{ground}}$  is simply as below:

$$R_{\text{ground}} = \text{IoU}(T_{\text{pred}}, T_{\text{gt}}) = \frac{|T_{\text{pred}} \cap T_{\text{gt}}|}{|T_{\text{pred}} \cup T_{\text{gt}}|} \quad (4)$$

**Object Counting.** For the counting task, our reward function provides a dense signal based on the average relative error across all shape categories. For each category  $k$ , we first compute a score  $R_{\text{count},k}$  that is inversely proportional to the relative error. Let  $y_k$  be the ground-truth count and  $\hat{y}_k$  be the predicted count for category  $k$ . The score for a single category is:

$$R_{\text{count},k} = \max\left(0, 1 - \frac{|\hat{y}_k - y_k|}{y_k + \varepsilon}\right) \quad (5)$$

Here, the absolute error is normalized by the magnitude of the ground-truth value, and a small constant  $\varepsilon$  (e.g.,  $10^{-9}$ ) ensures numerical stability. The final reward for the entire task,  $R_{\text{count}}$ , is the average of these scores over all  $K$  shape categories:

$$R_{\text{count}} = \frac{1}{K} \sum_{k=1}^K R_{\text{count},k} \quad (6)$$

**Temporal Jigsaw.** For the jigsaw puzzle, our reward function measures the structural correctness of the predicted sequence. We compute a penalty based on the cumulative displacement of elements from their correct positions. Let  $P_{\text{gt}}$  be the ground-truth permutation and  $\hat{P}$  be the predicted permutation. Let  $\text{pos}(v, P)$  denote the position of an element  $v$  in a sequence  $P$ . The total displacement error  $E_{\text{jigsaw}}$  is defined as:

$$E_{\text{jigsaw}} = \sum_{k=1}^n |\text{pos}(k, \hat{P}) - \text{pos}(k, P_{\text{gt}})| \quad (7)$$

This error is then normalized by the maximum possible error,  $E_{\text{max}}$ , which occurs for a reversed sequence. The final reward is given by:

$$R_{\text{jigsaw}} = 1 - \frac{E_{\text{jigsaw}}}{E_{\text{max}}} \quad (8)$$

## 4. Experiments

**Implementation Details.** Our VideoSSR-8B model is built upon the Qwen3-VL-8B-Instruct [39]. We perform RLVR on our newly constructed VideoSSR-30K dataset for one epoch. Key hyperparameters for training include a learning rate of  $1 \times 10^{-6}$ , a global batch size of 64, and a rollout number ( $N$ ) of 8 for generation, a KL divergence penalty with a coefficient of  $1 \times 10^{-3}$ . MAX\_FRAMES is configured to 48, and MAX\_PIXELS is set to  $256 \times 256$  for efficient training. The entire training process is conducted on 8 H200 GPUs and takes approximately 16 hours. To ensure a fair and reproducible comparison, both Qwen3-VL and VideoSSR are evaluated under identical conditions: FPS is set to 2, with MAX\_FRAMES configured to {32, 48, 64}. MAX\_PIXELS is set to  $512 \times 512$ . Greedy decoding is used to ensure reproducibility. Following recent research [55], chain of thought [52] is not utilized to mitigate hallucination [31] and ensure correct output formatting, therefore enhancing performance. The code is available at Supplementary Material.

**Benchmarks and Baselines.** To comprehensively evaluate the generalization capability of VideoSSR, we conduct experiments on 16 distinct benchmarks spanning four major video task categories:

- **General Video QA:** MVBench [25], TempCompass [29], AoTBench [58], and VinoGround [65].
- **Long Video QA:** Video-MME [13], LVBench [48], LongVideoBench [54], and CGBench [6].
- **Temporal Grounding:** QVHighlights [24], ActivityNet [5], CharadesSTA [14], and TACoS [40].

Table 3. Performance comparison on Temporal Grounding and Complex Reasoning tasks.

Model	Frames	Temporal Grounding				Complex Reasoning			
		QVHighlights	ActivityNet	CharadesSTA	TACoS	VideoMMMU	Video-TT	VCRBench	CVBench
<i>Closed Source Models</i>									
GPT-4o [34]	-	-	-	35.7	-	61.2	46.6	29.0	69.2
Gemini-1.5-Pro [45]	-	-	-	-	-	53.9	42.3	48.2	-
<i>Open Source Models</i>									
Qwen3-VL-8B-Instruct [39]	32	43.7	36.5	50.3	22.4	58.2	41.8	7.4	61.8
	48	46.4	38.4	50.0	25.9	58.5	43.0	7.4	61.5
	64	48.6	39.8	49.2	28.1	58.8	44.0	8.8	61.6
<b>VideoSSR-8B (Ours)</b>	32	59.6(+15.9)	42.1(+5.6)	52.1(+1.8)	23.1(+0.7)	59.9(+1.7)	44.2(+2.4)	10.7(+3.3)	63.5(+1.7)
	48	61.1(+14.7)	43.0(+4.6)	51.1(+1.1)	27.7(+1.8)	60.0(+1.5)	44.9(+1.9)	15.3(+7.9)	63.8(+2.3)
	64	62.6(+14.0)	43.7(+3.9)	49.9(+0.7)	30.6(+2.5)	60.9(+2.1)	45.8(+1.8)	17.8(+9.0)	63.3(+1.7)

• **Complex Reasoning:** VideoMMMU [21], Video-TT [68], VCRBench [41], and CVBench [70].

For the Temporal Grounding tasks, we report the Mean Intersection over Union (mIoU) as the primary evaluation metric. Further details regarding each benchmark and a full breakdown of the results can be found in Supplementary Material 1.1.

For our primary baseline, we select Qwen3-VL-8B-Instruct, as it represents the state-of-the-art among open-source models. To further contextualize the performance of our method, we also provide a comparative analysis against two formidable proprietary models: GPT-4o [34] and Gemini-1.5-Pro [45].

#### 4.1. Main Results

**General Video QA** As shown in the left half of Table 2, VideoSSR achieves substantial improvements on temporally related benchmarks such as VnoGround [65], even surpassing closed source models. It also obtains improvements on more general benchmarks, for instance on MVBench [25], achieving a score of 68.9 and similarly outperforming the closed source models.

**Long Video QA** As shown in the right half of Table 2, VideoSSR also achieves consistent improvements on four mainstream benchmarks. As we primarily conduct training and evaluation with a low number of frames, a gap remains compared to closed-source models on such long video understanding tasks, which is a direction for future research.

**Temporal Grounding** As shown in the left half of Table 3, benefiting from the Anomaly Grounding task, VideoSSR achieves remarkable zero-shot improvements on multiple mainstream temporal grounding benchmarks, especially on QVHighlights [24] and ActivityNet [5], with gains of +15.9 and +5.6, respectively.

**Complex Reasoning** As shown in the right half of Table 3, VideoSSR achieves a large improvement of +9.0 on VCRBench [41], a benchmark that is highly correlated with our

Temporal Jigsaw task. VideoSSR also obtains consistent improvements on other video reasoning benchmarks.

In summary, we validate the generalization capability of VideoSSR on the 16 aforementioned benchmarks. Notably, VideoSSR achieves consistent performance improvements across four major video tasks under three different frame settings. Under the 48 frame setting, VideoSSR obtains an average improvement of 5.1% across all 17 benchmarks (including VIUBench), comprehensively demonstrating the effectiveness of VideoSSR.

#### 4.2. Ablation Study

**Analysis on three pretext tasks.** First, we individually validate the effectiveness of the three pretext tasks, as shown in Table 4. Benefiting from its design, the Anomaly Grounding task leads to a significant performance increase on CharadesSTA. Similarly, the Temporal Jigsaw task brings a substantial boost to VCRBench. Notably, all three tasks individually improve performance on VideoMME, confirming their contribution to enhancing general video understanding capabilities.

Moreover, Table 4 also shows the impact of the smooth reward function on the results. We observe that the model trained with a strict matching reward function performs closer to the baseline. This is because a strict reward function often leads to sparse reward signals, which are more likely to result in a zero advantage in GRPO. Consequently, the training becomes inefficient, leading to smaller update magnitudes. Furthermore, strict reward function introduces training instability. For instance, training the anomaly grounding task with a strict reward function even degrades performance on CharadesSTA.

**Analysis on task diversity.** To further investigate the benefits of task diversity, we conduct a comparative analysis between single task training and our mixed task VideoSSR-30K, controlling for the data scale at 30k samples for both settings. As illustrated in Figure 5, we observe that simply

Table 4. **Ablation study of the three pretext tasks and their corresponding smooth reward functions.** G, C, and J represent Anomaly Grounding, Object Counting, and Temporal Jigsaw, respectively. ✓ indicates the component is used for training. R@0.5 denotes recall at an IoU threshold of 0.5. Step denotes step accuracy for VCRBench. The best and second best results are shown in bold and underlined.

Training Config				Understanding		Grounding		Reasoning	
Pretext Tasks			Reward	Video-MME	CharadesSTA	VCRBench			
G	C	J	Smooth	All	Long	mIoU	R@0.5	Acc	Step
<i>Baseline Model</i>									
✗	✗	✗	–	64.1	54.3	50.3	58.4	7.4	25.9
<i>Models on Subtasks</i>									
✓	✗	✗	✗	64.7	54.8	47.5	52.2	5.8	24.9
✓	✗	✗	✓	64.8	55.9	<b>53.8</b>	<b>63.8</b>	4.1	22.8
✗	✓	✗	✗	64.7	54.7	51.5	59.9	6.3	25.4
✗	✓	✗	✓	64.9	56.2	51.4	60.1	5.5	24.8
✗	✗	✓	✗	64.3	55.0	51.3	59.1	<u>13.4</u>	<u>32.7</u>
✗	✗	✓	✓	64.8	55.8	51.0	59.0	<b>15.9</b>	<b>35.5</b>
<i>Models on All Tasks</i>									
✓	✓	✓	✗	64.8	55.4	51.3	59.3	10.7	30.4
✓	✓	✓	✓	<b>65.2</b>	<b>57.1</b>	<u>52.1</u>	<u>60.6</u>	10.7	32.3

scaling up the data for a single task yields diminishing returns and even degrades performance. This finding suggests that designing a diverse set of pretext tasks, rather than focusing on a single one, is a more promising direction for enhancing model capabilities.

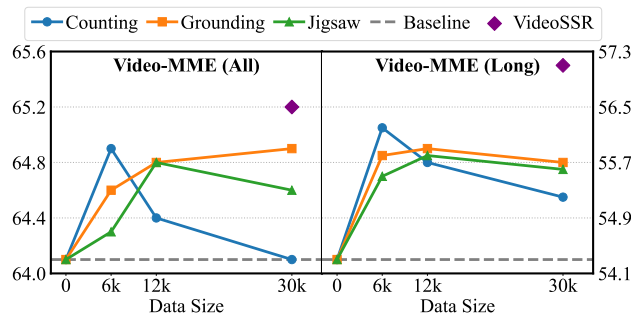


Figure 5. **Comparison of single task and mixed task training at the 30k data scale.** The results demonstrates that task diversity is more effective for improving performance than simply scaling up the data for a single pretext task.

We also compare our method against training with LongVideoReason [8] or ReWatch [64]. We utilize only the multiple-choice subsets from each dataset. The specific training procedures are:

- For **LongVideoReason**, the model is trained for 500 steps with a batch size of 64.
- For **ReWatch**, we use a composite subset of questions from its Video-R1 [11] and VideoEspresso [17] portions and train the model for one full epoch.

The results are presented in Table 5. Notably, the model trained with our VideoSSR-30K surpasses the performance of models trained on annotated datasets of a comparable

scale. Furthermore, we observe a critical limitation: fine-tuning the powerful Qwen3-VL on LongVideoReason, a dataset annotated by a less capable MLLM, can even lead to performance degradation, which further demonstrate the importance of our self-supervised paradigm.

Table 5. **Ablation study on different training datasets. “None” indicates the baseline Qwen3-VL.**

Training Config		Video-MME		CharadesSTA		VCRBench	
Fine-tuning Data	Size	All	Long	mIoU	R@0.5	Acc	Step
<i>Baseline Model</i>							
None	–	64.1	54.3	50.3	58.4	7.4	25.9
<i>Fine-tuned Models</i>							
LongVideoReason [8]	32k	63.6	53.3	51.7	59.4	7.1	26.1
ReWatch [64]	27k	64.7	56.7	51.6	59.2	2.7	22.2
<b>VideoSSR-30K</b>	<b>30K</b>	<b>65.2</b>	<b>57.1</b>	<b>52.1</b>	<b>60.6</b>	<b>10.7</b>	<b>32.3</b>

### Analysis on subtypes for the Anomaly Grounding task.

Finally, we investigate 14 distinct perturbation types and report their corresponding accuracies on Video-MME, as illustrated in Figure 6. Further details are provided in Supplementary Material 2.1. Based on these results, we select four perturbations that offer substantial improvements and create a uniform mixture to construct our final training set. Furthermore, we found that perturbations targeting temporal properties, such as simulating a fast forward effect by sampling denser frame sequences, does not appear to yield benefits and even introduced negative side effects. This may be because the base model, Qwen3-VL, relies on textual timestamps for its temporal awareness. Deliberately creating visual anomalies in this domain might confuse the model rather than enhance its learning.

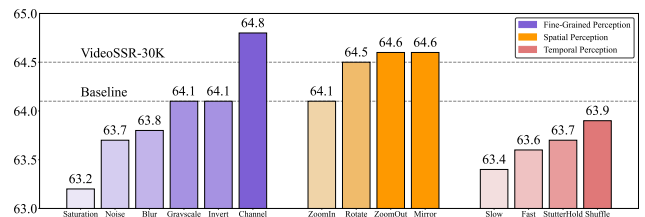


Figure 6. **Ablation study of the 14 perturbation subtypes for Anomaly Grounding.** Accuracy is reported on Video-MME.

## 5. Conclusion

In this paper, we introduce VideoSSR, a novel self-supervised reinforcement learning framework that addresses the critical data limitations in training MLLMs with novel pretext tasks. Our work establishes self-supervision as a powerful method for generating scalable, low-cost, high-quality training data, enabling models to learn directly from the intrinsic structure of video and move beyond the constraints of static, annotated datasets.

## References

- [1] Unaiza Ahsan, Rishi Madhok, and Irfan Essa. Video jigsaw: Unsupervised learning of spatiotemporal context for video action recognition. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 179–189. IEEE, 2019. 3
- [2] Lei Bai, Zhongrui Cai, Yuhang Cao, Maosong Cao, Weihang Cao, Chiyu Chen, Haojong Chen, Kai Chen, Pengcheng Chen, Ying Chen, et al. Intern-s1: A scientific multimodal foundation model. *arXiv preprint arXiv:2508.15763*, 2025. 1
- [3] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhao-hai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, Junyang Lin, et al. Qwen2.5-VL Technical Report, 2025. 1, 5
- [4] Ellis Brown, Arijit Ray, Ranjay Krishna, Ross Girshick, Rob Fergus, and Saining Xie. SIMS-V: Simulated instruction-tuning for spatial video understanding. *arXiv preprint arXiv:2511.04668*, 2025. 2
- [5] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 961–970, 2015. 6, 7
- [6] Guo Chen, Yicheng Liu, Yifei Huang, Yuping He, Baoqi Pei, Jilan Xu, Yali Wang, Tong Lu, and Limin Wang. Cg-bench: Clue-grounded question answering benchmark for long video understanding, 2024. 6
- [7] Yi Chen, Yuying Ge, Rui Wang, Yixiao Ge, Junhao Cheng, Ying Shan, and Xihui Liu. Grpo-care: Consistency-aware reinforcement learning for multimodal reasoning. *arXiv preprint arXiv:2506.16141*, 2025. 2
- [8] Yukang Chen, Wei Huang, Baifeng Shi, Qinghao Hu, Hanrong Ye, Ligeng Zhu, Zhijian Liu, Pavlo Molchanov, Jan Kautz, Xiaojuan Qi, et al. Scaling rl to long videos. *arXiv preprint arXiv:2507.07966*, 2025. 1, 2, 8
- [9] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasapat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blisstein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025. 1, 5
- [10] Lu Dong, Haiyu Zhang, Han Lin, Ziang Yan, Xiangyu Zeng, Hongjie Zhang, Yifei Huang, Yi Wang, Zhen-Hua Ling, Limin Wang, et al. Videotg-rl: Boosting video temporal grounding via curriculum reinforcement learning on reflected boundary annotations. *arXiv preprint arXiv:2510.23397*, 2025. 2
- [11] Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, Junfei Wu, Xiaoying Zhang, Benyou Wang, and Xiangyu Yue. Video-rl: Reinforcing video reasoning in mllms. *arXiv preprint arXiv:2503.21776*, 2025. 1, 2, 8
- [12] Basura Fernando, Hakan Bilen, Efstratios Gavves, and Stephen Gould. Self-supervised video representation learning with odd-one-out networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3636–3645, 2017. 2, 3
- [13] Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24108–24118, 2025. 6
- [14] Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. Tall: Temporal activity localization via language query. In *Proceedings of the IEEE international conference on computer vision*, pages 5267–5275, 2017. 6
- [15] Dong Guo, Faming Wu, Feida Zhu, Fuxing Leng, Guang Shi, Haobin Chen, Haoqi Fan, Jian Wang, Jianyu Jiang, Jiawei Wang, et al. Seed1. 5-vl technical report. *arXiv preprint arXiv:2505.07062*, 2025. 5
- [16] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025. 1, 2, 5
- [17] Songhao Han, Wei Huang, Hairong Shi, Le Zhuo, Xiu Su, Shifeng Zhang, Xu Zhou, Xiaojuan Qi, Yue Liao, and Si Liu. Videospresso: A large-scale chain-of-thought dataset for fine-grained video reasoning via core frame selection. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 26181–26191, 2025. 8
- [18] Zefeng He, Xiaoye Qu, Yafu Li, Siyuan Huang, Daizong Liu, and Yu Cheng. Framethinker: Learning to think with long videos via multi-turn frame spotlighting. *arXiv preprint arXiv:2509.24304*, 2025. 2
- [19] Zefeng He, Xiaoye Qu, Yafu Li, Tong Zhu, Siyuan Huang, and Yu Cheng. Diffthinker: Towards generative multimodal reasoning with diffusion models. *arXiv preprint arXiv:2512.24165*, 2025. 3
- [20] Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, et al. Glm-4.1 v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. *arXiv e-prints*, pages arXiv–2507, 2025. 5
- [21] Kairui Hu, Penghao Wu, Fanyi Pu, Wang Xiao, Yuanhan Zhang, Xiang Yue, Bo Li, and Ziwei Liu. Video-mmmu: Evaluating knowledge acquisition from multi-discipline professional videos. *arXiv preprint arXiv:2501.13826*, 2025. 7
- [22] Yuqi Huo, Mingyu Ding, Haoyu Lu, Zhiwu Lu, Tao Xiang, Ji-Rong Wen, Ziyuan Huang, Jianwen Jiang, Shiwei Zhang, Mingqian Tang, et al. Self-supervised video representation learning with constrained spatiotemporal jigsaw. 2021. 3
- [23] Longlong Jing and Yingli Tian. Self-supervised visual feature learning with deep neural networks: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 43(11):4037–4058, 2020. 3
- [24] Jie Lei, Tamara L Berg, and Mohit Bansal. Detecting moments and highlights in videos via natural language queries.

- Advances in Neural Information Processing Systems*, 34: 11846–11858, 2021. 6, 7
- [25] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22195–22206, 2024. 6, 7
- [26] Xinhao Li, Ziang Yan, Desen Meng, Lu Dong, Xiangyu Zeng, Yinan He, Yali Wang, Yu Qiao, Yi Wang, and Limin Wang. Videochat-r1: Enhancing spatio-temporal perception via reinforcement fine-tuning. *arXiv preprint arXiv:2504.06958*, 2025. 1, 2
- [27] Daizong Liu, Xiaoye Qu, Yinzheng Wang, Xing Di, Kai Zou, Yu Cheng, Zichuan Xu, and Pan Zhou. Unsupervised temporal video grounding with deep semantic clustering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1683–1691, 2022. 3
- [28] Daizong Liu, Xiang Fang, Xiaoye Qu, Jianfeng Dong, He Yan, Yang Yang, Pan Zhou, and Yu Cheng. Unsupervised domain adaptative temporal sentence localization with mutual information maximization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3567–3575, 2024. 3
- [29] Yuanxin Liu, Shicheng Li, Yi Liu, Yuxiang Wang, Shuhuai Ren, Lei Li, Sishuo Chen, Xu Sun, and Lu Hou. Tempcompass: Do video llms really understand videos? *arXiv preprint arXiv:2403.00476*, 2024. 6
- [30] Yuhong Liu, Beichen Zhang, Yuhang Zang, Yuhang Cao, Long Xing, Xiaoyi Dong, Haodong Duan, Dahua Lin, and Jiaqi Wang. Spatial-ssrl: Enhancing spatial understanding via self-supervised reinforcement learning. *arXiv preprint arXiv:2510.27606*, 2025. 2
- [31] Mi Luo, Zihui Xue, Alex Dimakis, and Kristen Grauman. When thinking drifts: Evidential grounding for robust video reasoning. *arXiv preprint arXiv:2510.06077*, 2025. 6
- [32] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European conference on computer vision*, pages 69–84. Springer, 2016. 2, 3
- [33] Mehdi Noroozi, Hamed Pirsiavash, and Paolo Favaro. Representation learning by learning to count. In *Proceedings of the IEEE international conference on computer vision*, pages 5898–5906, 2017. 2, 3
- [34] OpenAI. Hello GPT-4o. OpenAI Blog, 2024. Accessed: 2024-05-14. 1, 6, 7
- [35] OpenAI. Chatgpt, 2025. 1, 2, 5
- [36] Kun Ouyang, Yuanxin Liu, Haoning Wu, Yi Liu, Hao Zhou, Jie Zhou, Fandong Meng, and Xu Sun. Spacer: Reinforcing mllms in video spatial reasoning. *arXiv preprint arXiv:2504.01805*, 2025. 2
- [37] Viorica Patraucean, Lucas Smaira, Ankush Gupta, Adria Recasens, Larisa Markeeva, Dylan Banarse, Skanda Koppula, Mateusz Malinowski, Yi Yang, Carl Doersch, et al. Perception test: A diagnostic benchmark for multimodal video models. *Advances in Neural Information Processing Systems*, 36:42748–42761, 2023. 2
- [38] Xiaoye Qu, Pengwei Tang, Zhikang Zou, Yu Cheng, Jianfeng Dong, Pan Zhou, and Zichuan Xu. Fine-grained iterative attention network for temporal language localization in videos. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 4280–4288, 2020. 2
- [39] Qwen. Qwen3-vl, 2025. 1, 5, 6, 7
- [40] Michaela Regneri, Marcus Rohrbach, Dominikus Wetzels, Stefan Thater, Bernt Schiele, and Manfred Pinkal. Grounding action descriptions in videos. *Transactions of the Association for Computational Linguistics*, 1:25–36, 2013. 6
- [41] Pritam Sarkar and Ali Etemad. Vcrbench: Exploring long-form causal reasoning capabilities of large video language models. *arXiv preprint arXiv:2505.08455*, 2025. 7
- [42] Madeline C Schiappa, Yogesh S Rawat, and Mubarak Shah. Self-supervised learning for videos: A survey. *ACM Computing Surveys*, 55(13s):1–37, 2023. 3
- [43] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024. 1, 2, 5
- [44] Yunlong Tang, Jing Bi, Pinxin Liu, Zhenyu Pan, Zhangyun Tan, Qianxiang Shen, Jiani Liu, Hang Hua, Junjia Guo, Yunzhong Xiao, et al. Video-lmm post-training: A deep dive into video reasoning with large multimodal models. *arXiv preprint arXiv:2510.05034*, 2025. 1
- [45] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 1, 6, 7
- [46] Guodong Wang, Yunhong Wang, Jie Qin, Dongming Zhang, Xiuguo Bao, and Di Huang. Video anomaly detection by solving decoupled spatio-temporal jigsaw puzzles. In *European Conference on Computer Vision*, pages 494–511. Springer, 2022. 3
- [47] Jiangliu Wang, Jianbo Jiao, and Yun-Hui Liu. Self-supervised video representation learning by pace prediction. In *European conference on computer vision*, pages 504–521. Springer, 2020. 3
- [48] Weihang Wang, Zehai He, Wenyi Hong, Yean Cheng, Xiaohan Zhang, Ji Qi, Xiaotao Gu, Shiyu Huang, Bin Xu, Yuxiao Dong, et al. Lvbench: An extreme long video understanding benchmark. *arXiv preprint arXiv:2406.08035*, 2024. 6
- [49] Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. InternV3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025. 1, 5
- [50] Ye Wang, Ziheng Wang, Boshen Xu, Yang Du, Kejun Lin, Zihan Xiao, Zihao Yue, Jianzhong Ju, Liang Zhang, Dingyi Yang, et al. Time-r1: Post-training large vision language model for temporal video grounding. *arXiv preprint arXiv:2503.13377*, 2025. 2
- [51] Zifu Wang, Junyi Zhu, Bo Tang, Zhiyu Li, Feiyu Xiong, Jiaqian Yu, and Matthew B Blaschko. Jigsaw-r1: A study of rule-based visual reinforcement learning with jigsaw puzzles. *arXiv preprint arXiv:2505.23590*, 2025. 3

- [52] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022. 6
- [53] Bo Wu, Shoubin Yu, Zhenfang Chen, Joshua B Tenenbaum, and Chuang Gan. Star: A benchmark for situated reasoning in real-world videos. *arXiv preprint arXiv:2405.09711*, 2024. 2
- [54] Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. Longvideobench: A benchmark for long-context interleaved video-language understanding. *Advances in Neural Information Processing Systems*, 37:28828–28857, 2024. 6
- [55] Penghao Wu, Yushan Zhang, Haiwen Diao, Bo Li, Lewei Lu, and Ziwei Liu. Visual jigsaw post-training improves mllms. *arXiv preprint arXiv:2509.25190*, 2025. 3, 5, 6
- [56] Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. Next-qa: Next phase of question-answering to explaining temporal actions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9777–9786, 2021. 2
- [57] Dejing Xu, Jun Xiao, Zhou Zhao, Jian Shao, Di Xie, and Yueting Zhuang. Self-supervised spatiotemporal learning via video clip order prediction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10334–10343, 2019. 2, 3
- [58] Zihui Xue, Mi Luo, and Kristen Grauman. Seeing the arrow of time in large multimodal models. *arXiv preprint arXiv:2506.03340*, 2025. 6
- [59] Ziang Yan, Xinhao Li, Yanan He, Zhengrong Yue, Xiangyu Zeng, Yali Wang, Yu Qiao, Limin Wang, and Yi Wang. Videochat-r1. 5: Visual test-time scaling to reinforce multimodal reasoning by iterative perception. *arXiv preprint arXiv:2509.21100*, 2025. 2
- [60] Shusheng Yang, Jihan Yang, Pinzhi Huang, Ellis Brown, Zihao Yang, Yue Yu, Shengbang Tong, Zihan Zheng, Yifan Xu, Muhan Wang, Danhao Lu, Rob Fergus, Yann LeCun, Li Fei-Fei, and Saining Xie. Cambrian-s: Towards spatial supersensing in video. *arXiv preprint arXiv:2511.04670*, 2025. 2
- [61] Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B Tenenbaum. Clevrer: Collision events for video representation and reasoning. *arXiv preprint arXiv:1910.01442*, 2019. 2
- [62] Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gao-hong Liu, Lingjun Liu, et al. Dapo: An open-source llm reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*, 2025. 5
- [63] Yu Zeng, Wenxuan Huang, Shiting Huang, Xikun Bao, Yukun Qi, Yiming Zhao, Qiuchen Wang, Lin Chen, Zehui Chen, Huaian Chen, et al. Agentic jigsaw interaction learning for enhancing visual perception and reasoning in vision-language models. *arXiv preprint arXiv:2510.01304*, 2025. 3
- [64] Congzhi Zhang, Zhibin Wang, Yinchao Ma, Jiawei Peng, Yihan Wang, Qiang Zhou, Jun Song, and Bo Zheng. Rewatch-r1: Boosting complex video reasoning in large vision-language models through agentic data synthesis. *arXiv preprint arXiv:2509.23652*, 2025. 1, 3, 8
- [65] Jianrui Zhang, Mu Cai, and Yong Jae Lee. Vinoground: Scrutinizing Imms over dense temporal reasoning with short videos. *arXiv preprint arXiv:2410.02763*, 2024. 6, 7
- [66] Kaiyan Zhang, Yuxin Zuo, Bingxiang He, Youbang Sun, Runze Liu, Che Jiang, Yuchen Fan, Kai Tian, Guoli Jia, Pengfei Li, et al. A survey of reinforcement learning for large reasoning models. *arXiv preprint arXiv:2509.08827*, 2025. 1
- [67] Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Video instruction tuning with synthetic data. *arXiv preprint arXiv:2410.02713*, 2024. 2
- [68] Yuanhan Zhang, Yunice Chew, Yuhao Dong, Aria Leo, Bo Hu, and Ziwei Liu. Towards video thinking test: A holistic benchmark for advanced video reasoning and understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20626–20636, 2025. 7
- [69] Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, Rui Men, An Yang, et al. Group sequence policy optimization. *arXiv preprint arXiv:2507.18071*, 2025. 5
- [70] Nannan Zhu, Yonghao Dong, Teng Wang, Xueqian Li, Shengjun Deng, Yijia Wang, Zheng Hong, Tiantian Geng, Guo Niu, Hanyan Huang, et al. Cvbench: Evaluating cross-video synergies for complex multimodal understanding and reasoning. *arXiv preprint arXiv:2508.19542*, 2025. 7