

ViSS-R1: Self-Supervised Reinforcement Video Reasoning

Bo Fang¹ Yuxin Song² Haoyuan Sun^{2,3} Xinyao Zhang^{2,3}
 Qiangqiang Wu¹ Wenhao Wu⁴ Antoni B. Chan¹✉

¹ City University of Hong Kong, ² Baidu Inc., ³ Tsinghua University, ⁴ Amazon AGI

bofang6-c@my.cityu.edu.hk, abchan@cityu.edu.hk

Abstract

Complex video reasoning remains a significant challenge for Multimodal Large Language Models (MLLMs), as current R1-based methodologies often prioritize text-centric reasoning derived from text-based and image-based developments. In video tasks, such strategies frequently underutilize rich visual information, leading to potential shortcut learning and increased susceptibility to hallucination. To foster a more robust, visual-centric video understanding, we start by introducing a novel self-supervised reinforcement learning GRPO algorithm (Pretext-GRPO) within the standard R1 pipeline, in which positive rewards are assigned for correctly solving pretext tasks on transformed visual inputs, which makes the model to non-trivially process the visual information. Building on the effectiveness of Pretext-GRPO, we further propose the ViSS-R1 framework, which streamlines and integrates pretext-task-based self-supervised learning directly into the MLLM’s R1 post-training paradigm. Instead of relying solely on sparse visual cues, our framework compels models to reason about transformed visual input by simultaneously processing both pretext questions (concerning transformations) and true user queries. This necessitates identifying the applied transformation and reconstructing the original video to formulate accurate final answers. Comprehensive evaluations on six widely-used video reasoning and understanding benchmarks demonstrate the effectiveness and superiority of our Pretext-GRPO and ViSS-R1 for complex video reasoning. Codes and models are available at <https://github.com/bofang98/ViSS-R1>.

1. Introduction

Beyond fundamental video understanding, complex video reasoning presents new challenges in inferring objects, relationships, events and causality from video content [63, 67, 70, 76]. Following the success of post-training via Reinforcement Learning with Verifiable Reward (RLVR) [19, 30] in the large language model domain, recent research

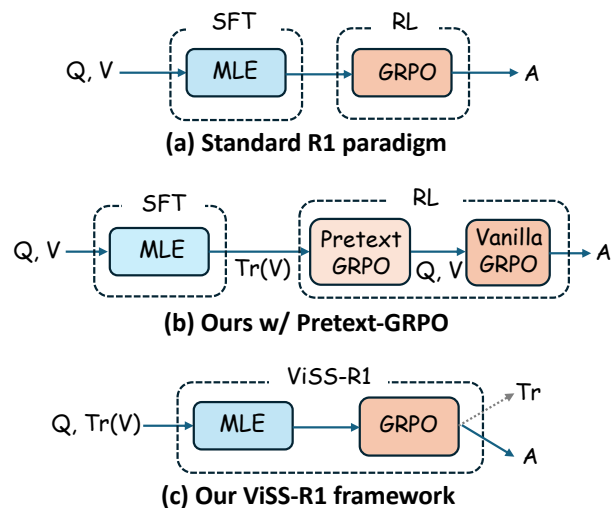


Figure 1. **Overview.** (a) Standard R1 paradigm for MLLM video reasoning consists of a SFT memorization stage followed by RL exploration. (b) We introduce an intermediate Pretext-GRPO stage for visual-centric RL reasoning by learning from self-supervised visual transformations $Tr(V)$. (c) Our ViSS-R1 framework fully integrates pretext-task reasoning into the R1-paradigm, where the model takes transformed visual inputs $Tr(V)$ and is designed to simultaneously output the inferred transformation (Tr) and the query’s answer (A).

has increasingly focused on incentivizing Multimodal Large Language Models (MLLMs) reasoning abilities for complex images [23, 40, 42, 50, 64, 71] and videos [14, 35, 48, 56], by revealing the thinking process in text-based Chain-of-Thought (CoT). Despite these early advances, video reasoning with MLLMs remains underexplored.

Inspired by image-based MLLM reasoning, recent video reasoning largely employs a similar framework based on the R1 paradigm [19, 53]. As illustrated in Fig. 1(a), MLLMs are first subjected to supervised finetuning (SFT) using CoTs, followed by refinement with reinforcement learning (RL) algorithms (e.g., PPO [46], GRPO [47]). To enhance temporal reasoning abilities, high-quality video-specific CoT datasets are carefully constructed for SFT

training. While for the RL optimization, methodological advances are limited to temporal augmentation [14] or temporal-aligned rewards addition [35, 56].

Within these existing video R1 works, dense video information is primarily used as contextual evidence, from which the model extracts sparse cues to support text-based reasoning. (See analysis in App. §A.) Despite the careful construction of Video CoT datasets, substantial video content remains underutilized, as videos typically contain much more information than can be conveyed by text. Consequently, models may exploit shortcuts by focusing on a single frame or snapshot, rather than engaging in deep, comprehensive video reasoning [68]. Moreover, the predominance of text-centric reasoning in current methodologies, which often overlooks rich visual information, increases the risk of hallucinated generations (Tab. 2). This raises an important question: how can we facilitate MLLM video reasoning from a *visual-centric* perspective?

We propose the **Video Self-Supervised Reinforcement** framework, termed as **ViSS-R1**. To address the challenge of sparse utilization of visual information in videos, we employ pretext-task-based self-supervised learning (SSL) to enhance visual-centric representation learning for the video. SSL has a rich history of leveraging various pretext transformations for representation learning, such as rotation [16], patch shuffling [11, 44], and inpainting [45] in images, as well as 3D rotation [28], clip shuffling [61], and acceleration [2, 65] in videos. The resulting pre-trained models are then transferred to fine-tune specific downstream tasks. Within the modern RLVR post-training framework, we utilize pretext SSL as an effective entry point for first extracting and understanding video information, before proceeding to targeted text-based and query-specific reasoning. This approach is motivated by the association that visual-centric pretext tasks provides natural reward signals well-suited for RL mechanism, without additional annotations.

In this paper, we propose two methods to introduce SSL into video-based MLLMs: 1) a warm-start pretext-GRPO algorithm based on SSL; 2) an integration of the pretext task into the full training pipeline of the MLLM. For the first approach, we introduce a separate Pretext-GRPO stage prior to standard RL training, aiming to provide a *warm start* for the RL policy model, in contrast to the cold start SFT (Fig. 1(b)). In Pretext-GRPO, visual inputs (including mixed images and videos) are subjected to various transformations, which are then used to train the policy model to identify the specific transformation applied. The optimization algorithm is identical to that of vanilla GRPO, except that the ground-truth labels correspond to specific transformation types. Subsequently, after applying normal GRPO optimization with real user questions, we demonstrate consistent improvement brought by Pretext-GRPO upon multiple video reasoning and understanding benchmarks.

In the above method, Pretext-GRPO is decoupled from the standard RL paradigm, which introduces additional training control complexity in practice. Thus, we propose to retain the standard R1 framework, but instead, compel MLLMs to directly reason about transformed videos (Fig. 1(c)). Specifically, the model is simultaneously prompted with both the pretext question and the user’s query, which necessitates the model to accurately answer the primary question by correctly identifying the applied transformation first, implicitly recovering the original videos. This integrated setting is maintained across both the SFT and RL stages: in SFT, the model learns to organize responses for both questions in a predefined format, while in RL, it is jointly optimized with two distinct reward signals. The complete pipeline constitutes our ViSS-R1.

In summary, our contributions are three-fold:

- We introduce a novel Pretext-GRPO RL algorithm for visual-centric MLLM video reasoning, utilizing self-supervised visual transformations. This annotation-free method can be seamlessly and effectively integrated into existing R1-paradigm frameworks.
- We establish the ViSS-R1 framework, which enhances video reasoning by training MLLMs to directly process transformed inputs in both SFT and RL stages.
- Comprehensive evaluations on multiple video reasoning and understanding benchmarks demonstrates the effectiveness and superiority of our approaches.

2. Related Work

2.1. Multimodal Large Language Model for Video

Video understanding is a fundamental task that aims to effectively manage, analyze, and interpret complicated video content [51]. Recent advances in large-scale visual instruction tuning led to powerful open-sourced MLLMs [1, 7, 32, 38, 55, 74], which have significantly improved video understanding. By focusing on video-specific spatio-temporal perception, current video MLLMs [3, 6, 8, 27, 36, 37, 62, 69, 72] have achieved notable progress in general tasks such as video question answering and captioning. Nevertheless, when addressing complex spatio-temporal video reasoning scenarios [22, 63, 75], these models remain inadequate due to a lack of thinking abilities.

Video reasoning presents new challenges in inferring objects, relationships and causality for visual spatial and temporal intelligence. Inspired by breakthroughs from OpenAI-o1 [25] and DeepSeek-R1 [19] in lifting the reasoning abilities of LLMs through RL, numerous studies [10, 23, 42, 49, 64, 71] have adopted the R1 paradigm to incentivize the visual reasoning abilities of MLLMs. Specifically for videos, Video-R1 [14] adapts the R1 pipeline to the video domain and introduces shuffling augmentation to improve temporal reasoning. VideoRFT [56] further incorporates a

Pretext Question

Please review the input video and determine whether the sequence of video frames has been reversed (i.e., played in reverse order).

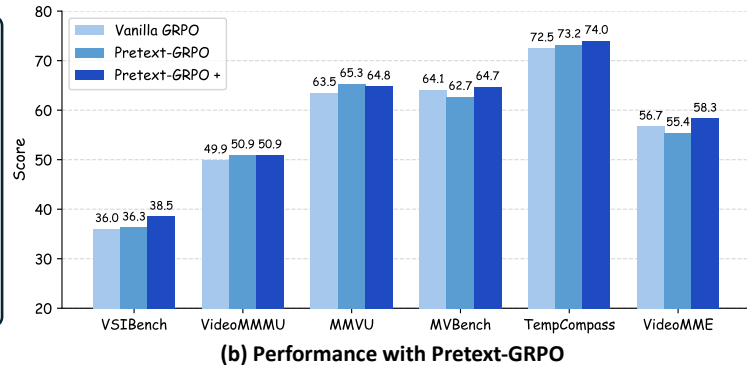
-The possible options are:

A. No, the video frames remain in their original order.

B. Yes, the video frames have been reversed.

Please focus on the temporal order of video contents and think about this question as if you were a human pondering deeply. Provide your detailed reasoning between the <think> </think> tags, and then give your final answer between the <answer> </answer> tags.

(a) "Reverse" MCQ example



(b) Performance with Pretext-GRPO

Figure 2. (a) Example of a "Reverse" MCQ pretext question used in our Pretext-GRPO, where randomly transformed visual inputs are leveraged to construct targeted pretext queries for policy model prompting. (b) Pretext-GRPO+ denotes a Pretext-GRPO stage followed by vanilla GRPO, which consistently improves performance across multiple video benchmarks. All results are based on 16-frame evaluation.

semantic-consistency reward to better align textual reasoning with visual information. Given the temporal structure and rich content of videos, researchers have focused on designing and integrating multiple types of temporal-related rewards [33, 35] into the original system. Despite these advances, current video reasoning approaches leverage only sparse visual cues for text-centric thinking, leaving substantial dense video information underutilized (§1). In contrast to previous works, we propose novel self-supervised RL algorithms that enable explicit visual processing, thereby enhancing subsequent deep video reasoning by utilizing more comprehensive video information.

2.2. Self-Supervised Representation Learning

Image-oriented self-supervised learning (SSL) has primarily employed transformation-based pretext tasks for upstream pretraining, such as solving jigsaws [5, 44], predicting rotation angles [16], localizing patches [11], etc., which have subsequently inspired analogous video SSL tasks like 3D rotation [28], space-time cubic puzzles [29]. Furthermore, various temporal-specific tasks, including sorting frames [31, 43] or clips [61], perceiving the arrow of time [59], discriminating among multiple temporal samplings [13, 26, 41], have been proposed for vision-only video models on extracting spatiotemporal representations for classification/recognition tasks. These effective designs motivate us to integrate such pretext transformations into the current RLVR framework for intermediate, visual-centric reasoning steps. Although contrastive learning algorithms [4, 17, 20] and reconstruction-based methods [21, 52, 54] have advanced SSL for large-scale vision foundation models, their complex training frameworks hinder direct adoption in standard RL settings.

Two concurrent works, Jigsaw-R1 [57] and VisualJigsaw [60], pursue similar self-supervised RL approaches by post-training MLLMs on the image jigsaw task. In contrast, we focus more on video reasoning and incorporate a

broader range of pretext transformations beyond single jigsaw objective. Furthermore, we systematically investigate a streamlined and effective integration of visual-centric SSL within the current RL framework, moving beyond simple combinations of pretext RL tasks.

3. Method

3.1. Vision-Oriented Pretext-GRPO

Preliminaries. Recent studies aiming to incentivize the reasoning capabilities of MLLMs typically adopt a paradigm consisting of SFT cold start followed by RL refinement. With carefully curated long CoT, the SFT stage enables the model to learn the semantics of specialized tag tokens and to autoregressively generate structured outputs through thinking before producing final answers, e.g., <think>...</think><answer>...</answer>.

During the RL stage, a commonly used algorithm is Group Relative Policy Optimization (GRPO) [47], which is a variant of Proximal Policy Optimization (PPO) [46], that eliminates the value model and significantly lowers the cost of training resources. Given an SFT initialized policy model π_θ and a reference model π_{ref} , GRPO first generates a set of responses $\{o_1, o_2, \dots, o_G\}$ through policy sampling. A group-based reward function then computes the corresponding rewards $\{r_1, r_2, \dots, r_G\}$ against the ground truth, which are subsequently used to estimate the advantage A_i for each response relative to the group:

$$A_i = \frac{r_i - \text{mean}(\{r_1, r_2, \dots, r_G\})}{\text{std}(\{r_1, r_2, \dots, r_G\})}. \quad (1)$$

Moreover, GRPO employs a clipped objective with a KL penalty term to stabilize the training process:

$$\begin{aligned} \mathcal{J}_{\text{GRPO}} = & \mathbb{E}_{q, \{o_i\}} \left[\frac{1}{G} \sum_{i=1}^G \left(\min \left(\frac{\pi_\theta(o_i|q)}{\pi_{\theta_{\text{old}}}(o_i|q)} A_i, \right. \right. \right. \\ & \left. \left. \text{clip} \left(\frac{\pi_\theta(o_i|q)}{\pi_{\theta_{\text{old}}}(o_i|q)}, 1 - \epsilon, 1 + \epsilon \right) A_i \right) - \right. \\ & \left. \beta \mathbb{D}_{\text{KL}}(\pi_\theta \parallel \pi_{\text{ref}}) \right]. \quad (2) \end{aligned}$$

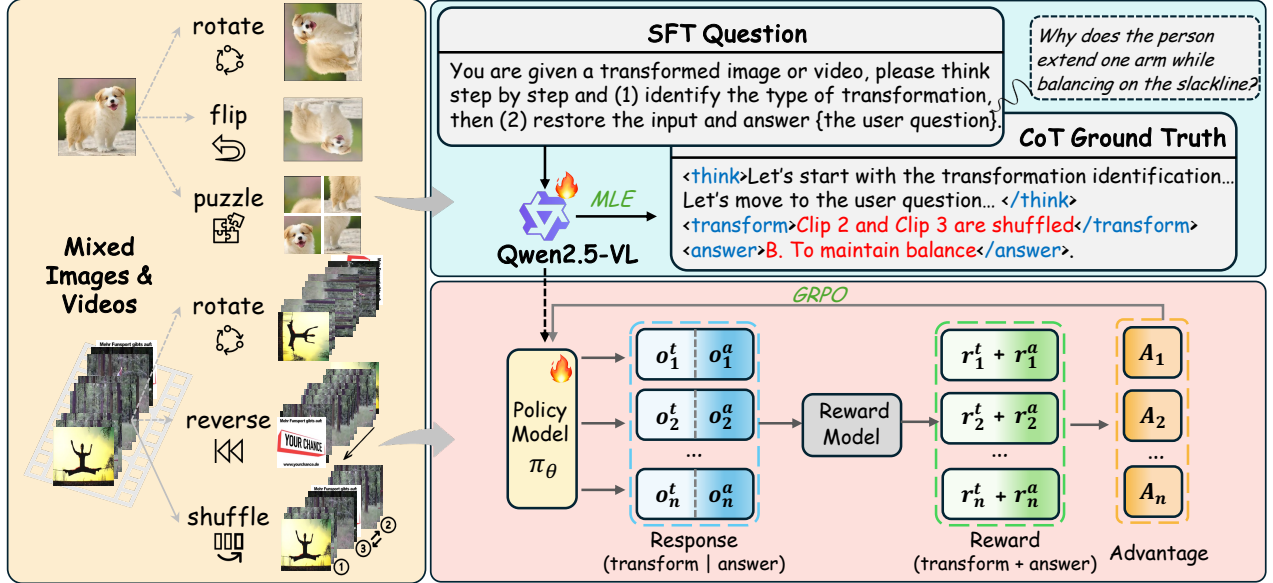


Figure 3. **ViSS-R1 framework.** Mixed images and videos are randomly augmented with SSL transformations for both SFT and RL reasoning. The models are required to simultaneously address pretext questions regarding the applied transformations and answer real user queries. ViSS-R1 additionally learns a `<transform>` tag in SFT to encapsulate pretext identification results, which provides structural organization for answers and facilitates reward manipulation during RL exploration.

Pretext-GRPO. Beyond trivial memorization in SFT [9], GRPO encourages models to reason creatively along diverse trajectories by rewarding those that lead to correct final outcomes. However, despite its advantages, vanilla GRPO in MLLMs often overlooks rich visual information when answering localized questions, particularly in videos [14, 56] (see §1). To address this limitation, we propose Pretext-GRPO that conducts RL on self-supervised transformations, enabling the policy model to effectively leverage and explore rich, vision-oriented information.

Given a training image or video V , we do not use its real user question Q for Pretext-GRPO. Instead, a handcrafted pretext question Q_p , together with the corresponding transformed visual input $\text{Tr}(V)$, are inputs to the policy model for RL. To facilitate straightforward reward assignment, all Q_p are formulated as multiple-choice questions (MCQs). Specifically for images, we consider three types of self-supervised transformations (SSTs), enabling the model to learn spatial semantics by solving spatial-related tasks:

- *Rotate* [16]: The image is randomly rotated by 0° , 90° , 180° , or 270° , resulting in a 4-option MCQ.
- *Flip*: The image is randomly flipped vertically or horizontally, including no flip, forming a 3-option MCQ.
- *Puzzle* [11, 44]: The image is evenly divided into four patches by splitting it along both the horizontal and vertical axes; two patches are then randomly selected and their positions swapped, resulting in a 6-option MCQ.

Similarly, three SSTs are randomly applied to videos to capture spatio-temporal representations:

- *3D Rotate* [28]: All frames in a video are randomly ro-

tated by the same degree, resulting in a 4-option MCQ.

- *Reverse* [59]: The video is either presented in its original direction or reversed, forming a binary MCQ.
- *Shuffle* [61]: The video is divided into 4 consecutive clips, of which 2 are shuffled, yielding a 6-option MCQ.

We present an example of a “Reverse” pretext question in Fig. 2(a), with full prompts of these questions in App. §B.

The above SSTs with MCQ formulations provide “free” supervision that can be interpreted as rewards during RL. As in vanilla GRPO, our Pretext-GRPO assigns a reward of 1 if the pretext question is correctly answered, and 0 otherwise. The policy model π_θ is then updated using the GRPO objective in (2), based on the pretext question Q_p and transformed input $\text{Tr}(V)$, i.e., $\frac{\pi_\theta(o_i|\text{Tr}(V), Q_p)}{\pi_{\theta_{\text{old}}}(o_i|\text{Tr}(V), Q_p)}$. Unlike cold-start SFT, which employs token-level regression to memorize and imitate the reasoning path [9], Pretext-GRPO provides a *warm start* for the policy model by enabling it to self-examine internal visual content and semantics prior to standard RL. The effectiveness of Pretext-GRPO is demonstrated in Fig. 2(b) across multiple video reasoning and understanding benchmarks (more in §4.1). When combining vanilla GRPO on true user questions and raw visual input after Pretext-GRPO (denoted as **Pretext-GRPO+**), we observe consistent improvements across all benchmarks.

3.2. Vision-Integrated ViSS-R1

Initiating RL with a separate Pretext-GRPO stage helps stabilize standard RL training by warm-starting the policy model to recognize spatiotemporal visual transforma-

tions. However, splitting RL into two distinct stages introduces additional training control complexity because the two stages optimize different objectives. This motivates us to integrate SSL into a single-stage RL procedure that directly answers user questions based on transformed image/video inputs. We therefore introduce the ViSS-R1 framework (see Fig. 3).

SFT with transform-tagged CoT. Base MLLMs are generally unable to answer two questions concurrently in a structured manner, even with explicit instructions. To enable the categorization of two answers from single-round outputs, we teach models to learn generating responses in a predefined structure using SFT with transform-tagged CoTs. Building on previous CoT construction, we introduce an additional `<transform></transform>` tag to encapsulate the results of the final predicted transformation. Following Video-R1 [14], our ground-truth CoT rationales are reconstructed by distilling answers from the advanced Qwen2.5-VL-72B model [1] on randomly transformed images and videos (Sec. 3.1). During SFT, we apply the *identical* transformation to each input as used for prompting the 72B model and train the base model using the MLE loss, as illustrated in Fig. 3 (top right). This process enables the base model to learn to reason and respond to both questions within distinct tags, which correspond to the two rewards in the subsequent RL stage.

RL reasoning on transformed inputs. Given transformed visual inputs $\text{Tr}(V)$, a pretext question Q_p and a real user query Q , our transformation-acquainted SFT model is required to solve the self-supervised task (implicitly restoring the inputs) and answering real questions via GRPO algorithm. Compared to vanilla GRPO, the key distinction of ViSS-R1 is that two types of responses (o_i^t for transformation identification result and o_i^q for user question answers) are sampled in a single generation round. The importance sampling ratio between the new policy π_θ and the old policy $\pi_{\theta_{\text{old}}}$ in (2) can be rewritten as:

$$\frac{\pi_\theta(o_i^t, o_i^q | \text{Tr}(V), Q_p, Q)}{\pi_{\theta_{\text{old}}}(o_i^t, o_i^q | \text{Tr}(V), Q_p, Q)}. \quad (3)$$

Reward manipulation. ViSS-R1 incorporates three types of rewards for RL: ① **Transformation reward** R_t : The initial task is to identify the transformation applied to images or videos, and a reward of 0.5 is assigned if the pretext question Q_p is correctly answered, and 0 otherwise. ② **Accuracy reward** R_a : Following previous works [14, 56], we adopt task-specific accuracy metrics, including Exact Match for multiple-choice and numerical questions, ROUGE for open-ended generation, Word Error Rate for OCR tasks, and a scaled relative accuracy for regression problems. ③ **Format reward** R_f : The model’s output is required to follow a predefined structure: the thinking process must be enclosed in

`<think>...</think>` tags, transformation prediction in `<transform>...</transform>`, and answers to user questions in `<answer>...</answer>`. The overall reward for a transformed sample is $R = R_t + R_a + R_f$.

4. Experiment

4.1. Experimental Setup

Benchmarks and metric. Following recent work [14, 56, 73], we evaluate our model on three general video reasoning benchmarks including VSI-Bench [63], VideoMMU [22], MMVU [75], and three general video understanding benchmarks: MVBench [34], TempCompass [39] and VideoMME [15] (w/o subtitles). For MMVU, only the multiple-choice subset are used for evaluation. We report average accuracy (Acc) for all above tasks.

Model training. We use Qwen2.5-VL-7B [1] as the base model, and our training dataset is derived from Video-R1 [14] (i.e., Video-R1-CoT-165k and Video-R1-260k). For efficiency, we sample 32 frames from each video and limit the maximum resolution of each frame to $128 \times 28 \times 28$ during training. Pretext-GRPO is trained for 500 steps, followed by vanilla GRPO for 1K steps, denoted as Pretext-GRPO+. In integrated ViSS-R1, we reprompt Qwen2.5-VL-72B for transformation-acquainted CoT construction and SFT training. All models are trained on 8 NVIDIA A800 (80G) GPUs. Our codebase is built on Open-R1 [12].

Inference. For inference, we sample 32 frames and increase the resolution to $256 \times 28 \times 28$ to enhance performance, following works [14, 56]. The decoding configuration follows the Qwen2.5-VL demo, with top_p=0.001 and temperature=0.01. During ViSS-R1 inference, pretext questions are removed from the testing prompt, and models are required to reason on untransformed, raw videos.

4.2. Comparisons with Previous Methods

General video performance. We present comprehensive comparisons with previous methods in Tab. 1. Notably, our proposed methods, **Pretext-GRPO+** and **ViSS-R1**, achieve state-of-the-art performance on 4 out of 6 video reasoning and understanding benchmarks (**39.2%** on VSI-Bench, **53.9%** on VideoMMU, **75.3%** on TempCompass, and **60.5%** on VideoMME). These improvements demonstrate the effectiveness of our self-supervised RL strategy in incentivizing the video reasoning capabilities of MLLMs. Additionally, compared to the initial baseline Qwen2.5-VL, ViSS-R1 achieves consistent and significant improvements of **+7.2%** on VSI-Bench, **+4.3%** on VideoMMU, and **+6.1%** on MMVU on reasoning-specific benchmarks, highlighting the value of post-training techniques for unlocking video reasoning abilities. Relative to another baseline, Video-R1 (since we use identical image and video data), Pretext-GRPO+ yields **+3.4%** on VSI-Bench and **+1.6%**

Table 1. **Performance comparisons on general video benchmarks.** VSI refers to VSI-Bench, and TempC refers to the TempCompass benchmark. * indicates results reported with higher input resolutions and more frames (i.e., 768×28×28 and 768 sampled frames), while our results use 256×28×28 and 32 frames. **Boldface** and underline indicate the best and second-best results, respectively.

Model	Venue	Video Reasoning			Video Understanding		
		VSI.	VideoMMMU	MMVU	MVBench	TempC.	VideoMME
<i>Proprietary Models</i>							
GPT-4o [24]		34.0	61.2	75.4	-	-	71.9
<i>Open-Source Models</i>							
LLaMA-VID [36]	ECCV 24	-	-	-	41.9	45.6	-
ShareGPT4Video [3]	NeurIPS 24	-	-	-	51.2	-	39.9
VideoLLaMA2 [8]	arXiv 24.06	-	-	44.8	54.6	-	47.9
LongVA-7B [72]	TMLR 24	29.2	23.9	-	-	56.9	52.6
VILA-1.5-8B [37]	CVPR 24	28.9	20.8	-	-	58.8	-
Video-UTR-7B [68]	ICLR 25	-	-	-	58.8	59.7	52.6
mPLUG-Owl3-8B [66]	ICLR 25	-	-	-	54.5	-	53.5
LLaVA-OneVision-7B [32]	TMLR 24	32.4	33.8	49.2	56.7	-	58.2
Qwen2.5-VL-7B [1]	arXiv 25.02	30.1	48.1	60.0	59.0	72.6	56.6
<i>R1-based Models</i>							
Video-R1 [14]	NeurIPS 25	35.8	52.3	63.8	63.9	73.2	59.3
TinyLLaVA-Video-R1 [73]	arXiv 25.04	-	-	46.9	-	49.5	46.6
VideoChat-R1 [35]	arXiv 25.04	-	-	-	67.9*	-	-
Temporal-RLT [33]	arXiv 25.06	-	-	65.0	68.1	73.3	57.6
VideoRFT [56]	NeurIPS 25	36.8	51.1	68.5	62.1	73.7	59.8
Video-KTR [58]	ICLR 26	-	<u>52.6</u>	65.9	-	73.4	<u>60.3</u>
Pretext-GRPO+ (Ours)		39.2	53.9	65.3	<u>66.0</u>	<u>73.9</u>	60.1
ViSS-R1 (Ours)		<u>37.3</u>	51.7	<u>66.1</u>	65.6	75.3	60.5

Table 2. **Comparisons on the HallusionBench.**

Model	HallusionBench		
	aAcc	fAcc	qAcc
Qwen2.5-VL-7B [1]	61.8	35.0	36.0
Video-R1 [14]	58.8 (-3.0)	32.9 (-2.1)	33.2 (-2.8)
VideoRFT [56]	59.8 (-2.0)	35.0 (-0.0)	32.7 (-3.3)
ViSS-R1 (Ours)	61.4 (-0.4)	37.6 (+2.6)	38.0 (+2.0)

on VideoMMMU, while ViSS-R1 achieves **+2.1%** on TempCompass and **+1.2%** on VideoMME. The superior performance demonstrates the strong potential of integrating self-supervised RL into existing R1-based methods for visual-centric complement.

Hallucination analysis. Tab. 2, reports the performance of various models on HallusionBench [18]. As discussed in §1, purely outcome-driven RL models [14, 56] suffer from severe hallucinations. In contrast, our SSL-RL mechanism (ViSS-R1) substantially mitigates this issue.

4.3. Ablation Study

Training data. ViSS-R1 is trained on mixed images and videos from Video-R1, utilizing 6 types of self-supervised transformations. We do not ablate each transformation individually, as all pretext tasks have been verified effective in priors (§2.2), and several have already been successfully incorporated into current R1 paradigms [57, 60]. In Tab. 3, we divide the training sources into *Image only* (A.1) and *Video only* (A.2) to evaluate the effectiveness of image-based (ro-

tate, flip, and puzzle) and video-based (rotate, reverse, and shuffle) transformations, separately. Results indicate that image data, combined with 2D pretext tasks, achieves reasonable performance on most video benchmarks, providing sufficient spatial knowledge for basic video understanding. Meanwhile, video-specific SSR1 (A.2) consistently outperforms its image-based counterpart, underscoring the critical role of spatiotemporal representation learning for video domain reasoning and understanding.

Training paradigm. We first evaluate the effectiveness of Pretext-GRPO as an intermediate and independent RL stage in Fig. 2(b). Although not using real questions, Pretext-GRPO alone achieves notable results on reasoning benchmarks such as VideoMMMU (50.9%) and MMVU (65.3%), supporting our primary motivation of acquiring vision-centric knowledge prior to localized answering. Pretext-GRPO+, which integrates vanilla GRPO with user questions, further enhances and stabilizes performance on more general video understanding tasks.

Moreover, we analyze the sequential SFT and RL training paradigm in ViSS-R1 as shown in Tab. 3. *SFT only* (B.1) refers to training exclusively with supervised fine-tuning on our transformation-reprompt CoT annotations. In this stage, the model learns to follow specific instructions and produce structured responses; however, it lacks evident reasoning abilities, and may suffer from the overfitting problem. In contrast, *RL only* (B.2), despite omitting SFT initialization,

Table 3. **Ablation studies of ViSS-R1.** Our full model is jointly trained on mixed images and videos, employing a sequential SFT and RL training pipeline. RL supervision is provided via three types of rewards (transformation, accuracy, and format rewards), combined as $R = R_f + R_t + R_a$. VSI refers to VSI-Bench, and TempC refers to the TempCompass benchmark. **Boldface** indicate the best results.

Config	Model	Video Reasoning			Video Understanding		
		VSI.	VideoMMMU	MMVU	MVBench	TempC.	VideoMME
<i>Training Data</i>							
A.1	Image only	34.8	49.1	65.6	63.9	75.0	57.8
A.2	Video only	37.0	51.2	65.3	64.7	75.2	59.1
<i>Training Paradigm</i>							
B.1	SFT only	31.7	46.4	63.0	60.9	70.6	54.7
B.2	RL only	24.0	51.4	66.6	65.6	73.5	59.3
<i>Reward Modeling</i>							
C.1	$R = R_f + R_t$	32.0	48.7	64.5	62.3	74.3	56.0
C.2	$R = R_f + R_a$	35.8	51.6	63.2	63.6	73.9	58.0
<i>Full Model</i>							
D	ViSS-R1	37.3	51.7	66.1	65.6	75.3	60.5

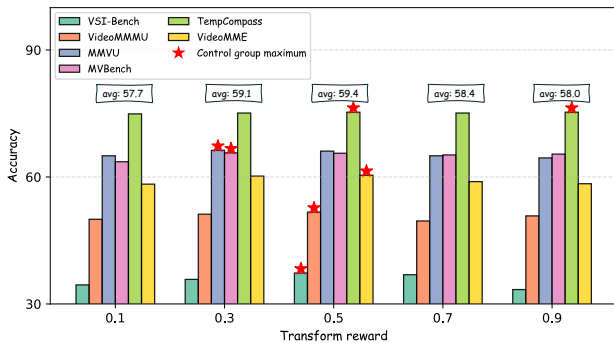


Figure 4. **Impact of training reward R_t .**

demonstrates strong reasoning capabilities on most benchmarks. Notably, the RL-only model achieves significantly lower performance (24.0%) on VSI-Bench, which contains many challenging regression-type tasks. We observe that the RL-only model struggles to generate well-formatted predictions for such tasks, underscoring the necessity of an SFT cold start. By adopting standard R1 paradigm (SFT followed by RL), our ViSS-R1 achieves robust and generalizable improvements.

Reward modeling. In Tab. 3 (C.1, C.2), we ablate the impact of reward design, with format reward R_f fixed as an anchor following previous literature [14, 56]. As demonstrated in Pretext-GRPO, reinforcement learning with only self-supervised transformation identification can yield solid improvements over its SFT model (C.1 vs. B.1), particularly on MMVU (64.5%) and TempCompass (74.3%). Furthermore, when jointly trained with accuracy reward R_a on real questions (D), we achieve a comprehensively enhanced and superior model, ViSS-R1.

Scaling transformation reward R_t . Intuitively, pretext tasks should be assigned less importance than real questions when jointly optimized in ViSS-R1. Here, we analyze the impact of transformation reward R_t while keeping the correctly answered accuracy reward R_a as 1.0 (Fig. 4). Across

Table 4. **Comparisons on transformed videos.** V denotes raw videos, while $Tr(V)$ represents transformed videos with rotation, reversal, or shuffling. Performance difference Δ is also reported.

Model	VSI-Bench			VideoMME		
	V	$Tr(V)$	Δ	V	$Tr(V)$	Δ
Video-R1 [14]	35.8	32.6	-3.2	59.3	55.9	-3.4
VideoRFT [56]	36.8	30.8	-6.0	59.8	54.6	-5.2
Pretext-GRPO+ (Ours)	39.2	38.1	-1.1	60.1	58.3	-1.8
ViSS-R1 (Ours)	37.3	34.8	-2.5	60.5	59.3	-1.2

all control group experiments, setting $R_t = 0.5$ yields the highest average score of 59.4% and achieves the best results in 4 out of 6 video benchmarks. Increasing R_t (e.g., to 0.9) leads to a clearly drop in performance, as the model receives a similar reward for answering pretext and real questions, which diminishes its ability to distinguish between them and ultimately impairs its capability.

Comparisons on transformed videos. We evaluate various models on *transformed* videos using both pretext and real questions in the testing prompts. The real question accuracy is presented in Tab. 4. As expected, prior R1 models (Video-R1 and VideoRFT) exhibit significant performance drops due to perturbed video sequences and domain-shift. Our models demonstrate strong robustness to diverse video augmentations, with minimal accuracy loss. Finally, for our models, accuracy on the untransformed videos (V) is better than on transformed videos, which demonstrates the efficacy of the pretext task – processing of the transformed videos in the pretext task encourages useful aggregation of information that transfers to untransformed videos.

4.4. Training Curves

We monitor key configuration dynamics during RL in Fig. 6. The transformation reward R_t in ViSS-R1 increases rapidly within the first 100 iterations and then stabilizes, indicating that the model quickly acquires pretext reasoning ability from the SFT stage. Compared to vanilla GRPO only



Question: If I am standing by the fireplace and facing the tv, is the stool to my left, right, or back? An object is to my back if I would have to turn at least 135 degrees in order to face it. A. back B. right C. left VSI/Bench/arkit/scenes/45260928.mp4

<think>Let me think about this for a moment. If I am standing by the fireplace and facing the TV, I need to determine the position of the stool relative to me. The question specifies that an object is to my back if I would have to turn at least 135 degrees to face it. First, let's break it down. If I am facing the TV, I need to consider the direction of the stool. If I turn 135 degrees to face the stool, it would mean the stool is not directly in front of me, but rather to the side or behind me. Now, let's consider the options:
 A. back - This would mean the stool is directly behind me, which would require a 180-degree turn to face it.
 B. right - This would mean the stool is to my right, which would require a 90-degree turn to face it.
 C. left - This would mean the stool is to my left, which would require a 90-degree turn to face it.
 Hmm, if I need to turn at least 135 degrees to face the stool, it must be further back than just to the right or left. It must be directly behind me, which is the furthest direction from the TV.
 Oh, I see. The stool is indeed to my back, as it would require a 135-degree turn to face it. This aligns with the definition given in the question. Let's verify this one last time. If I turn 135 degrees to face the stool, it is not just to the right or left, but directly behind me. This confirms that the stool is to my back. Therefore, the correct answer is A.</think>
 <answer>A</answer>

Video-R1's reasoning

Directly answer question and compare options



<think>The video shows a cozy room with various elements like a fireplace, a TV, and a stool. Let's analyze the scene step by step:
 1. First, we see the wooden floor leading into the room. The room has a fireplace with a mirror above it and a TV placed on a stand next to the fireplace. There is a stool near the fireplace.
 2. The room is decorated with paintings on the walls and a chandelier hanging from the ceiling. The floor is wooden, and the room is well-lit with natural light coming through the windows.
 3. The stool is located near the fireplace, and it appears to be positioned in front of the fireplace when facing the TV. If someone is standing by the fireplace and facing the TV, the stool is directly in front of them.
 4. To determine the relative position of the stool, we need to consider the orientation of the person and the objects around them. If the person is facing the TV, the stool is to their left side.
 5. The stool is placed on the floor, and there is no indication that the person would have to turn more than 90 degrees to face the stool. The person can easily see the stool while facing the TV.
 6. The stool is not to the right or back of the person; it is to their left. If it were to the right or back, the person would have to turn at least 135 degrees to face it.
 7. The person can maintain a natural posture while facing the TV and the stool is within their field of view.
 Therefore, if the person is standing by the fireplace and facing the TV, the stool is to their left.</think>
 <answer>C</answer>

ViSS-R1's reasoning

Analyze video



Answer question



Filter options

Figure 5. Qualitative results. For ViSS-R1, we remove the pretext question in prompts and use untransformed videos for inference.

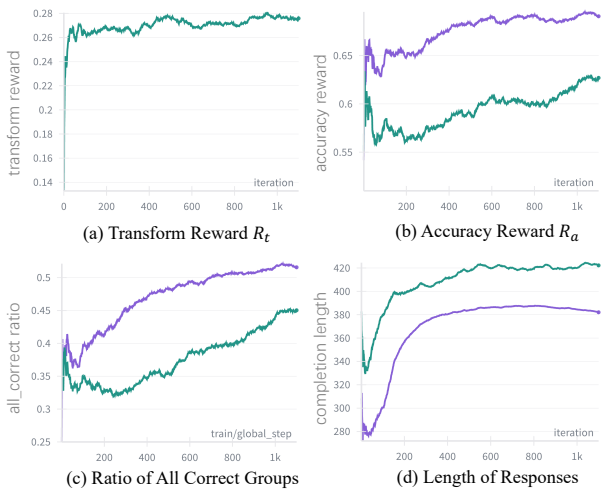


Figure 6. Training curves. Purple lines represent results from vanilla GRPO, while green lines correspond to our ViSS-R1.

on user questions, our accuracy reward R_a is noticeably lower (Fig. 6(b)), due to the increased difficulty introduced by visual transformations. Meanwhile, as task difficulty increases, the all correct ratio (the proportion of groups in which all generations are correct) decreases compared to vanilla GRPO. A higher all correct ratio suggests that more training samples are not truly learned (advantage is 0), which also explains why ViSS-R1 could outperform previous methods. Furthermore, ViSS-R1 produces a longer average response (Fig. 6(d)), as it addresses two questions in

a single forward pass.

4.5. Qualitative Result

We show qualitative results in Fig. 5. Previous Video-R1 typically attempts to solve the problem directly and analyze the options (in MCQs), often without carefully reviewing the referenced video. However, visual reasoning is essential for correct inference in this context. In contrast, our ViSS-R1, even without explicit instructions, exhibits a pattern of first analyzing the video content, then answering the question, and finally verifying the answer by checking available options. The problem-solving solution aligns well with our initial objective: enabling visual-centric reasoning. More examples can be found in App. §C.

5. Conclusion

We propose Pretext-GRPO, a pretext-based self-supervised reinforcement learning approach, along with its integration into the R1-style framework, ViSS-R1, to address visual-centric complex video reasoning. By leveraging annotation-free SSL transformations, Pretext-GRPO enables thorough examination of visual content prior to text-based answering. The integrated ViSS-R1 further streamlines the training process and consistently improves overall video reasoning performance. We believe incorporating SSL mechanisms into MLLMs represents a promising direction for the development of future intelligent multimodal models.

References

- [1] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 2, 5, 6
- [2] Sagie Benaim, Ariel Ephrat, Oran Lang, Inbar Mosseri, William T Freeman, Michael Rubinstein, Michal Irani, and Tali Dekel. Speednet: Learning the speediness in videos. In *CVPR*, pages 9922–9931, 2020. 2
- [3] Lin Chen, Xilin Wei, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Zhenyu Tang, Li Yuan, et al. Sharegpt4video: Improving video understanding and generation with better captions. *NeurIPS*, 37:19472–19495, 2024. 2, 6
- [4] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, pages 1597–1607. Pmlr, 2020. 3
- [5] Yingyi Chen, Xi Shen, Yahui Liu, Qinghua Tao, and Johan AK Suykens. Jigsaw-vit: Learning jigsaw puzzles in vision transformer. *Pattern Recognition Letters*, 166:53–60, 2023. 3
- [6] Yukang Chen, Fuzhao Xue, Dacheng Li, Qinghao Hu, Ligeng Zhu, Xiuyu Li, Yunhao Fang, Haotian Tang, Shang Yang, Zhijian Liu, et al. Longvila: Scaling long-context visual language models for long videos. *arXiv preprint arXiv:2408.10188*, 2024. 2
- [7] Zhe Chen, Jiannan Wu, Wenhao Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *CVPR*, pages 24185–24198, 2024. 2
- [8] Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, et al. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv preprint arXiv:2406.07476*, 2024. 2, 6
- [9] Tianzhe Chu, Yuexiang Zhai, Jihan Yang, Shengbang Tong, Saining Xie, Dale Schuurmans, Quoc V Le, Sergey Levine, and Yi Ma. Sft memorizes, rl generalizes: A comparative study of foundation model post-training. In *ICML*, 2025. 4
- [10] Yihe Deng, Hritik Bansal, Fan Yin, Nanyun Peng, Wei Wang, and Kai-Wei Chang. Openvlthinker: An early exploration to complex vision-language reasoning via iterative self-improvement. *arXiv preprint arXiv:2503.17352*, 2025. 2
- [11] Carl Doersch, Abhinav Gupta, and Alexei A Efros. Unsupervised visual representation learning by context prediction. In *ICCV*, pages 1422–1430, 2015. 2, 3, 4
- [12] Hugging Face. Open r1: A fully open reproduction of deepseek-r1, january 2025. URL <https://github.com/huggingface/open-r1>, page 9, 2025. 5
- [13] Bo Fang, Wenhao Wu, Chang Liu, Yu Zhou, Dongliang He, and Weipin Wang. Mamico: Macro-to-micro semantic correspondence for self-supervised video representation learning. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 1348–1357, 2022. 3
- [14] Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, Junfei Wu, Xiaoying Zhang, Benyou Wang, and Xiangyu Yue. Video-r1: Reinforcing video reasoning in mllms. *arXiv preprint arXiv:2503.21776*, 2025. 1, 2, 4, 5, 6, 7
- [15] Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *CVPR*, pages 24108–24118, 2025. 5
- [16] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. *arXiv preprint arXiv:1803.07728*, 2018. 2, 3, 4
- [17] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent—a new approach to self-supervised learning. *NeurIPS*, 33:21271–21284, 2020. 3
- [18] Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, et al. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *CVPR*, pages 14375–14385, 2024. 6
- [19] 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
- [20] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, pages 9729–9738, 2020. 3
- [21] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, pages 16000–16009, 2022. 3
- [22] 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. 2, 5
- [23] Wenxuan Huang, Bohan Jia, Zijie Zhai, Shaosheng Cao, Zheyu Ye, Fei Zhao, Zhe Xu, Yao Hu, and Shaohui Lin. Vision-r1: Incentivizing reasoning capability in multimodal large language models. *arXiv preprint arXiv:2503.06749*, 2025. 1, 2
- [24] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 6
- [25] Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. Openai o1 system card. *arXiv preprint arXiv:2412.16720*, 2024. 2
- [26] Simon Jenni, Givi Meishvili, and Paolo Favaro. Video representation learning by recognizing temporal transformations. In *ECCV*, pages 425–442. Springer, 2020. 3

- [27] Peng Jin, Ryuichi Takanobu, Wancai Zhang, Xiaochun Cao, and Li Yuan. Chat-univi: Unified visual representation empowers large language models with image and video understanding. In *CVPR*, pages 13700–13710, 2024. 2
- [28] Longlong Jing, Xiaodong Yang, Jingen Liu, and Yingli Tian. Self-supervised spatiotemporal feature learning via video rotation prediction. *arXiv preprint arXiv:1811.11387*, 2018. 2, 3, 4
- [29] Dahun Kim, Donghyeon Cho, and In So Kweon. Self-supervised video representation learning with space-time cubic puzzles. In *AAAI*, pages 8545–8552, 2019. 3
- [30] Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, et al. Tulu 3: Pushing frontiers in open language model post-training. *arXiv preprint arXiv:2411.15124*, 2024. 1
- [31] Hsin-Ying Lee, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. Unsupervised representation learning by sorting sequences. In *ICCV*, pages 667–676, 2017. 3
- [32] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. 2, 6
- [33] Hongyu Li, Songhao Han, Yue Liao, Junfeng Luo, Jialin Gao, Shuicheng Yan, and Si Liu. Reinforcement learning tuning for videollms: Reward design and data efficiency. *arXiv preprint arXiv:2506.01908*, 2025. 3, 6
- [34] 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 *CVPR*, pages 22195–22206, 2024. 5
- [35] 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, 3, 6
- [36] Yanwei Li, Chengyao Wang, and Jiayi Jia. Llama-vid: An image is worth 2 tokens in large language models. In *ECCV*, pages 323–340. Springer, 2024. 2, 6
- [37] Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. Vila: On pre-training for visual language models. In *CVPR*, pages 26689–26699, 2024. 2, 6
- [38] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023. 2
- [39] 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. 5
- [40] Ziyu Liu, Zeyi Sun, Yuhang Zang, Xiaoyi Dong, Yuhang Cao, Haodong Duan, Dahua Lin, and Jiaqi Wang. Visual-rft: Visual reinforcement fine-tuning. *arXiv preprint arXiv:2503.01785*, 2025. 1
- [41] Dezhao Luo, Chang Liu, Yu Zhou, Dongbao Yang, Can Ma, Qixiang Ye, and Weiping Wang. Video cloze procedure for self-supervised spatio-temporal learning. In *AAAI*, pages 11701–11708, 2020. 3
- [42] Fanqing Meng, Lingxiao Du, Zongkai Liu, Zhixiang Zhou, Quanfeng Lu, Daocheng Fu, Tiancheng Han, Botian Shi, Wenhai Wang, Junjun He, et al. Mm-eureka: Exploring the frontiers of multimodal reasoning with rule-based reinforcement learning. *arXiv preprint arXiv:2503.07365*, 2025. 1, 2
- [43] Ishan Misra, C Lawrence Zitnick, and Martial Hebert. Shuffle and learn: unsupervised learning using temporal order verification. In *ECCV*, pages 527–544. Springer, 2016. 3
- [44] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *ECCV*, pages 69–84. Springer, 2016. 2, 3, 4
- [45] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *CVPR*, pages 2536–2544, 2016. 2
- [46] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017. 1, 3
- [47] 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, 3
- [48] Yudi Shi, Shangzhe Di, Qirui Chen, and Weidi Xie. Enhancing video-llm reasoning via agent-of-thoughts distillation. In *CVPR*, pages 8523–8533, 2025. 1
- [49] Haoyuan Sun, Jiaqi Wu, Bo Xia, Yifu Luo, Yifei Zhao, Kai Qin, Xufei Lv, Tiantian Zhang, Yongzhe Chang, and Xueqian Wang. Reinforcement fine-tuning powers reasoning capability of multimodal large language models. *arXiv preprint arXiv:2505.18536*, 2025. 2
- [50] Huajie Tan, Yuheng Ji, Xiaoshuai Hao, Minglan Lin, Pengwei Wang, Zhongyuan Wang, and Shanghang Zhang. Reason-rft: Reinforcement fine-tuning for visual reasoning. *arXiv preprint arXiv:2503.20752*, 2025. 1
- [51] Yunlong Tang, Jing Bi, Siting Xu, Luchuan Song, Susan Liang, Teng Wang, Daoan Zhang, Jie An, Jingyang Lin, Rongyi Zhu, et al. Video understanding with large language models: A survey. *TCSVT*, 2025. 2
- [52] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *NeurIPS*, 35:10078–10093, 2022. 3
- [53] Luong Trung, Xinbo Zhang, Zhanming Jie, Peng Sun, Xiaoran Jin, and Hang Li. Reft: Reasoning with reinforced fine-tuning. In *ACL (Volume 1: Long Papers)*, pages 7601–7614, 2024. 1
- [54] Limin Wang, Bingkun Huang, Zhiyu Zhao, Zhan Tong, Yinan He, Yi Wang, Yali Wang, and Yu Qiao. Videomae v2: Scaling video masked autoencoders with dual masking. In *CVPR*, pages 14549–14560, 2023. 3
- [55] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 2

- [56] Qi Wang, Yanrui Yu, Ye Yuan, Rui Mao, and Tianfei Zhou. Videorf: Incentivizing video reasoning capability in mllms via reinforced fine-tuning. *arXiv preprint arXiv:2505.12434*, 2025. 1, 2, 4, 5, 6, 7
- [57] 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, 6
- [58] Ziyue Wang, Sheng Jin, Zhongrong Zuo, Jiawei Wu, Han Qiu, Qi She, Hao Zhang, and Xudong Jiang. Video-ctr: Reinforcing video reasoning via key token attribution. *arXiv preprint arXiv:2601.19686*, 2026. 6
- [59] Donglai Wei, Joseph J Lim, Andrew Zisserman, and William T Freeman. Learning and using the arrow of time. In *CVPR*, pages 8052–8060, 2018. 3, 4
- [60] 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, 6
- [61] Dejing Xu, Jun Xiao, Zhou Zhao, Jian Shao, Di Xie, and Yueting Zhuang. Self-supervised spatiotemporal learning via video clip order prediction. In *CVPR*, pages 10334–10343, 2019. 2, 3, 4
- [62] Lin Xu, Yilin Zhao, Daquan Zhou, Zhijie Lin, See Kiong Ng, and Jiashi Feng. Pllava: Parameter-free llava extension from images to videos for video dense captioning. *arXiv preprint arXiv:2404.16994*, 2024. 2
- [63] Jihan Yang, Shusheng Yang, Anjali W Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in space: How multimodal large language models see, remember, and recall spaces. In *CVPR*, pages 10632–10643, 2025. 1, 2, 5
- [64] Yi Yang, Xiaoxuan He, Hongkun Pan, Xiyan Jiang, Yan Deng, Xingtao Yang, Haoyu Lu, Dacheng Yin, Fengyun Rao, Minfeng Zhu, et al. R1-onevision: Advancing generalized multimodal reasoning through cross-modal formalization. *arXiv preprint arXiv:2503.10615*, 2025. 1, 2
- [65] Yuan Yao, Chang Liu, Dezhao Luo, Yu Zhou, and Qixiang Ye. Video playback rate perception for self-supervised spatio-temporal representation learning. In *CVPR*, pages 6548–6557, 2020. 2
- [66] Jiabo Ye, Haiyang Xu, Haowei Liu, Anwen Hu, Ming Yan, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. mplug-owl3: Towards long image-sequence understanding in multi-modal large language models. *arXiv preprint arXiv:2408.04840*, 2024. 6
- [67] Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B Tenenbaum. Clevrer: Collision events for video representation and reasoning. In *ICLR*, 2020. 1
- [68] En Yu, Kangheng Lin, Liang Zhao, Yana Wei, Zining Zhu, Haoran Wei, Jianjian Sun, Zheng Ge, Xiangyu Zhang, Jingyu Wang, et al. Unhackable temporal rewarding for scalable video mllms. *arXiv preprint arXiv:2502.12081*, 2025. 2, 6
- [69] Sicheng Yu, Chengkai Jin, Huanyu Wang, Zhenghao Chen, Sheng Jin, Zhongrong Zuo, Xiaolei Xu, Zhenbang Sun, Bingni Zhang, Jiawei Wu, et al. Frame-voyager: Learning to query frames for video large language models. *arXiv preprint arXiv:2410.03226*, 2024. 2
- [70] Haoji Zhang, Xin Gu, Jiawen Li, Chixiang Ma, Sule Bai, Chubin Zhang, Bowen Zhang, Zhichao Zhou, Dongliang He, and Yansong Tang. Thinking with videos: Multimodal tool-augmented reinforcement learning for long video reasoning. *arXiv preprint arXiv:2508.04416*, 2025. 1
- [71] Jingyi Zhang, Jiaxing Huang, Huanjin Yao, Shunyu Liu, Xikun Zhang, Shijian Lu, and Dacheng Tao. R1-vl: Learning to reason with multimodal large language models via step-wise group relative policy optimization. *arXiv preprint arXiv:2503.12937*, 2025. 1, 2
- [72] Peiyuan Zhang, Kaichen Zhang, Bo Li, Guangtao Zeng, Jingkang Yang, Yuanhan Zhang, Ziyue Wang, Haoran Tan, Chunyuan Li, and Ziwei Liu. Long context transfer from language to vision. *arXiv preprint arXiv:2406.16852*, 2024. 2, 6
- [73] Xingjian Zhang, Siwei Wen, Wenjun Wu, and Lei Huang. Tinyllava-video-r1: Towards smaller llms for video reasoning. *arXiv preprint arXiv:2504.09641*, 2025. 5, 6
- [74] 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
- [75] Yilun Zhao, Haowei Zhang, Lujing Xie, Tongyan Hu, Guo Gan, Yitao Long, Zhiyuan Hu, Weiyuan Chen, Chuhan Li, Zhijian Xu, et al. Mmvu: Measuring expert-level multi-discipline video understanding. In *CVPR*, pages 8475–8489, 2025. 2, 5
- [76] Bolei Zhou, Alex Andonian, Aude Oliva, and Antonio Torralba. Temporal relational reasoning in videos. In *ECCV*, pages 803–818, 2018. 1