VisionCoach: Reinforcing Grounded Video Reasoning via Visual-Perception Prompting

Daeun Lee, Shoubin Yu, Yue Zhang, and Mohit Bansal

University of North Carolina, Chapel Hill, USA
{daeun,shoubin,yuezhan,mbansal}@cs.unc.edu
https://visioncoach.github.io/

Abstract. Video reasoning requires models to locate and track question-relevant evidence across frames. While reinforcement learning (RL) with verifiable rewards improves accuracy, it still struggles to achieve reliable spatio-temporal grounding during the reasoning process. Moreover, improving grounding typically relies on scaled training data or inference-time perception tools, which increases annotation cost or computational cost. To address this challenge, we propose VisionCoach, an input-adaptive RL framework that improves spatio-temporal grounding through visual prompting as training-time guidance. During RL training, visual prompts are selectively applied to challenging inputs to amplify question-relevant evidence and suppress distractors. The model then internalizes these improvements through self-distillation, enabling grounded reasoning directly on raw videos without visual prompting at inference. VisionCoach consists of two components: (1) Visual Prompt Selector, which predicts appropriate prompt types conditioned on the video and question, and (2) Spatio-Temporal Reasoner, optimized with RL under visual prompt guidance and object-aware grounding rewards that enforce object identity consistency and multi-region bounding-box IoU. Extensive experiments demonstrate that VisionCoach achieves state-of-the-art performance across diverse video reasoning, video understanding, and temporal grounding benchmarks (V-STAR, VideoMME, World-Sense, VideoMMMU, PerceptionTest, and Charades-STA), while maintaining a single efficient inference pathway without external tools. Our results highlight the effectiveness of training-time visual prompting as a lightweight mechanism for improving grounded video reasoning.

Keywords: Video Reasoning · Visual Prompting · Reinforcement Learning · Self-Distillation

1 Introduction

Recent advances in reinforcement learning (RL) with verifiable rewards [36] have begun to improve multimodal reasoning and are increasingly applied to video reasoning [6, 12, 15, 27]. Despite recent progress, existing video reasoning approaches still struggle to achieve reliable spatio-temporal grounding throughout the reasoning process. Text-centric video reasoning models [14, 29, 45] (Fig. 1 (a))

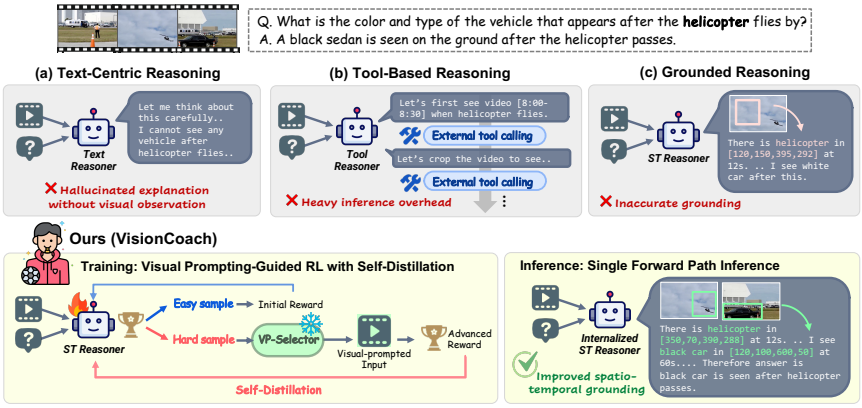


Fig. 1: Comparison with previous video reasoning approaches and VisionCoach. Comparing with other video reasoning models, VisionCoach leverages visual-prompt-guided RL with self-distillation to internalize improved spatio-temporal grounding behaviors induced by visual guidance. At inference time, it maintains a single forward-pass reasoning while achieving enhanced grounding performance.

often generate hallucinated explanations driven by language priors rather than faithful visual observations. Visual tool-calling approaches [18, 38, 41, 50, 53, 56] (Fig. 1 (b)) improve grounding by invoking external perception tools such as temporal clipping or zoom-in. While these tools can retrieve relevant evidence, the reasoning model itself does not internalize grounded perception and remains dependent on external modules at inference time. Recent grounded reasoning models [38] (Fig. 1 (c)) attempt to integrate spatio-temporal evidence within a single model by interleaving grounding and reasoning. Nevertheless, grounding remains unreliable, frequently producing inaccurate object references or hallucinated bounding boxes that propagate errors during reasoning.

At the core of this issue is the absence of mechanisms that enforce alignment between intermediate reasoning steps and spatio-temporal evidence. In practice, improving grounding typically requires either scaling training data or additional perception modules (e.g., cropping image regions or trimming the video) at inference time. However, dense annotations across diverse video datasets are expensive to obtain, while inference-time tools increase computational overhead. These solutions improve grounding only through heavier external intervention, rather than enhancing the model's intrinsic perception behavior. Instead, we shift the focus from data scaling and inference-time intervention to **training-time guidance**, aiming to internalize grounded reasoning behavior while preserving a lightweight inference pipeline.

To this end, we propose VisionCoach, an input-adaptive RL framework that uses visual prompts as a training-time vision coach to improve spatio-temporal grounding while enabling inference directly on raw videos. The key idea is to selectively apply visual prompts to challenging inputs during RL training to strengthen

grounding, and to internalize these improvements through self-distillation so that visual prompting is no longer required at inference. Instead of relying solely on implicit feature-space attention, VisionCoach performs reasoning-driven perception control at the input level, amplifying question-relevant evidence and suppressing distractors during training.

Specifically, VisionCoach consists of two components: (i) Visual Prompt Selector (VP-Selector), which predicts an appropriate visual prompt type conditioned on the video and question to provide adaptive perception guidance for challenging inputs, and (ii) Spatio-Temporal Reasoner (ST-Reasoner), which is optimized with RL under visual prompt-guided perception and grounding-aware rewards. Before training ST-Reasoner, we first construct a visual prompting candidate dataset using a proxy reasoner and train VP-Selector with SFT. During ST-Reasoner training, hard samples are identified based on initial rewards, and the frozen VP-Selector predicts visual prompt types that guide the ST-Reasoner to attend to relevant regions and moments, producing more grounded reasoning trajectories. To further strengthen grounded reasoning, we introduce an object-aware spatial grounding reward that enforces object identity consistency and average IoU across multiple predicted bounding boxes, encouraging accurate multi-object grounding and temporally consistent reasoning. To remove prompt dependency at inference, we employ self-distillation, where the prompted input serves as a coach to guide the policy model, enabling grounded perception behavior to be internalized. At inference, the model performs reasoning directly on raw videos with a single forward pass, without visual prompting. Together, these designs provide localized perception guidance during training while keeping a simple and efficient inference pipeline.

We evaluate VisionCoach on the (1) spatio-temporal reasoning benchmark V-STAR [7], (2) general video understanding benchmarks (VideoMME [15], WorldSense [25], VideoMMMU [27], and PerceptionTest [40]), and (3) temporal grounding benchmark (Charades-STA [17]). On V-STAR, VisionCoach surpasses GPT-4o and improves over Qwen2.5-VL-7B by +15.0% mAM and +25.1% mLGM, establishing new state-of-the-art performance. In general video understanding and temporal grounding benchmarks, it consistently outperforms prior open-source VideoLLMs, demonstrating strong performance in long-video reasoning and perception-oriented understanding tasks. We further provide comprehensive analyses, including spatio-temporal attention map visualization, component-wise ablation, generalization effect of VP-Selector across different backbone models, and statistics of adaptive visual prompting.

Our contributions are summarized as follows:

- We propose an input-adaptive RL framework for video reasoning that explicitly guides spatio-temporal grounding through training-time visual prompting and self-distillation, enabling the reasoning model to internalize grounded perception without requiring visual prompts at inference time.
- We design an object-aware spatial grounding reward that incorporates object identity consistency and multi-region bounding-box IoU to better support spatial-grounded reasoning.

- We introduce a visual prompt selector with a proxy-reasoner-based data construction pipeline to predict appropriate visual prompts for video QA inputs.
- Extensive experiments and analyses demonstrate that VisionCoach achieves state-of-the-art performance across a wide range of video reasoning, video understanding, and temporal grounding benchmarks.

2 Related Work

Video Reasoning. Video understanding has advanced rapidly with large multimodal models [1, 31, 33], enabling complex video QA and reasoning across diverse areas [7, 11, 15, 25, 27, 40, 47, 52]. Despite progress, reliable *grounded* video reasoning remains challenging because models must track object states, events, and interactions over time. A growing line of work applies reinforcement learning with verifiable or rule-based rewards to strengthen multimodal reasoning [4, 8, 21, 35, 38, 46, 50, 53, 56], but existing approaches often either (i) remain text-centric and hallucinate evidence, or (ii) rely on iterative tool operations at inference time (e.g., progressive ROI localization), introducing overhead and leaving limited explicit control over dense spatio-temporal evidence. Our work targets this gap by using *training-time* perception control to improve grounding.
Visual Prompting. Visual prompting [48] augments inputs with lightweight visual cues (e.g., boxes, masks, points, scribbles, overlays) to steer attention and improve localized understanding without heavy architectural changes. Early and concurrent studies show that simple edits in pixel space, such as drawing a red circle around an object, can reliably direct VLM attention [43]. Recent work expands visual prompting to general vision understanding in MLLMs [3, 9, 20, 32]. For video settings, prompting has also been used to improve temporal grounding by adding structured cues such as per-frame numbering [49, 57]. Meanwhile, learned or automated prompting methods aim to *select* or *retrieve* effective prompts conditioned on the input, improving robustness and usability [58, 59]. In VisionCoach, we use adaptive prompting *only during RL training* and then internalize the benefit so inference no longer depends on prompting.
Model Distillation. Knowledge distillation transfers behavior from a stronger teacher to a student model, improving efficiency and robustness [24]. Beyond classical teacher-student training, self-distillation and iterative teacher replacement can further improve generalization without additional labels [16]. Distillation has also been used in sequential decision making (policy distillation) to compress or unify behaviors learned via RL [42]. In VisionCoach, self-distillation helps the ST-Reasoner *internalize* the grounding improvements induced by visual prompting guided training trajectories, enabling a single, prompt-free inference.

3 Method

As shown in Fig. 2, we propose VisionCoach, an input-adaptive RL framework that improves spatio-temporal grounding in video reasoning through visual

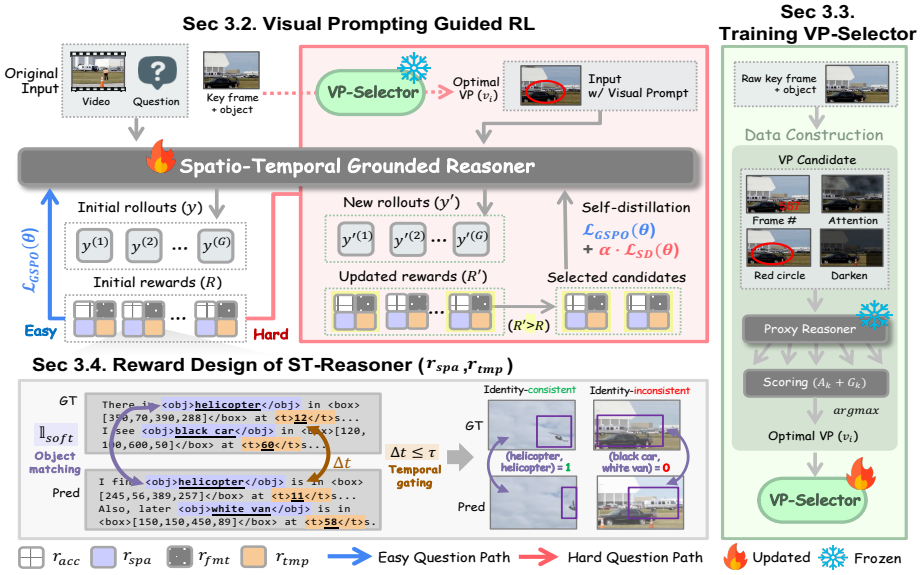


Fig. 2: Detailed Architecture of VisionCoach. We introduce VisionCoach, a visual-prompt-guided RL framework for training a spatio-temporally grounded reasoner (Sec. 3.2). The framework includes Visual Prompt Selector (VP-Selector) that predicts optimal visual prompts for challenging inputs (Sec. 3.3), and object-aware spatial grounding rewards that enforce object identity consistency and multiple predicted bounding boxes (Sec. 3.4).

prompting guidance. We begin by formalizing spatio-temporal grounded reasoning and presenting the motivation for our approach (Sec. 3.1). Next, we introduce the overall training pipeline of VisionCoach in Sec. 3.2, which integrates visual prompting with RL and self-distillation to encourage grounded reasoning. Finally, we describe the key components in detail: the Visual Prompt Selector (Sec. 3.3), which predicts input-adaptive visual prompts, followed by the reward design of Spatio-Temporal Reasoner (Sec. 3.4), which learns grounded reasoning from prompt-guided training signals and grounding-aware rewards.

3.1 Problem Statement and Motivation

Given an input video x and a question q, the goal of video QA is to generate a grounded reasoning trajectory y and predict the final answer a over complex and dynamic visual content. Unlike text-based reasoning (Fig. 1 (a)), our reasoning requires a policy model π_θ to integrate explicit spatio-temporal grounding, identifying when and where relevant evidence occurs while reasoning.

Motivation. To better understand the relationship between grounding and answering performance, we analyze model behaviors on the PerceptionTest subset. As shown in Fig. 3, correctly answered samples consistently exhibit higher

temporal alignment, object identity matching, and spatial IoU compared to incorrectly answered ones, indicating that accurate spatio-temporal grounding is correlated with correct answering. We further investigate how different visual prompts affect answering performance. As shown in Tab. 1, different prompts lead to different results, suggesting that selecting an appropriate visual prompt is crucial for effective answering [49,57,59]. Notably, the oracle prompt selection achieves significantly higher accuracy, indicating that adaptive perceptual guidance can substantially improve downstream answering when the appropriate prompt is applied. Motivated by these, we propose an RL framework that enables input-adaptive visual prompting to achieve reliable grounding and answering.

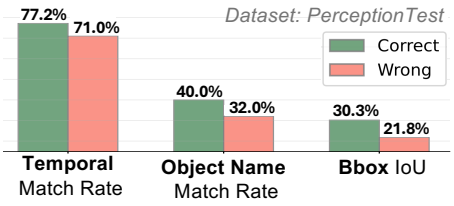


Fig. 3: Grounding effect on answering.

Models	Raw key frame	Darken	Red circle	Oracle (Any-corr)
Qwen2.5-VL	52.2	43.3	51.5	**70.8**
Gemini-2.5-Flash	47.4	50.4	49.3	**75.2**

Table 1: Effect of visual prompting.

3.2 VisionCoach: VP-Guided RL with Self-Distillation

We introduce VisionCoach, an RL framework that integrates VP-Selector and ST-Reasoner to enable input-adaptive visual guidance during training. We describe each component in detail in Sec. 3.3 (VP-Selector) and Sec. 3.4 (Reward design for ST-Reasoner). The ST-Reasoner π_θ is optimized using GSPO [61], starting from a cold-start initialization and trained on video–question pairs (x, q). For each input, G reasoning trajectories are sampled and evaluated with grounding-aware rewards. Visual prompts are selectively applied to challenging inputs during RL to strengthen spatio-temporal grounding, and their benefits are internalized through self-distillation, eliminating the need for visual prompting at inference. The overall training procedure is shown in Algorithm 1.

Input-adaptive hard sample identification. Given an input (x_i, q_i), we first perform G initial rollouts using the current policy. Let $\{R_i^{(g)}\}_{g=1}^{G}$ denote the resulting overall rewards. We compute the average reward $\bar{R}_i$ and classify the sample as hard if $\bar{R}_i < k$, where k is a predefined threshold for hard sample filtering. This input-adaptive mechanism determines whether additional visual guidance is necessary for each example, allowing more targeted visual prompting to help with grounding.

Visual prompting guidance generation. For hard samples, we feed (x_i, q_i) into the trained VP-Selector, which is elaborated in Sec. 3.3 to obtain the optimal visual prompt (such as darken), which is denoted as v_i. We then apply this prompt to the key frames of x_i to construct a visual-prompted input x_i', and append a textual `hint` describing the applied visual prompting to the original question q_i and make q_i'. The visual prompt introduces localized cues that

Algorithm 1 VisionCoach: Visual Prompt Guided RL with Self-Distillation

Require: Dataset $\mathcal{D} = \{(x_i, q_i)\}_{i=1}^{N}$, ST-Reasoner π_θ, trained VP-Selector, reward r, number of rollouts G, hard question threshold k, self-distillation weight α.

for each $(x_i, q_i) \in \mathcal{D}$ **do**
  for $g \leftarrow 1$ to G **do**
    $y_i^{(g)} \sim \pi_\theta(\cdot \mid x_i, q_i),\ R_i^{(g)} \leftarrow r(x_i, y_i^{(g)})$ `//* Conduct initial rollout`
  end for
  $\bar{R}_i \leftarrow \frac{1}{G}\sum_{g=1}^{G} R_i^{(g)},\ \mathbb{I}_i \leftarrow \mathbf{1}[\bar{R}_i < k]$
  if $\mathbb{I}_i = 1$ **then** `//* Generate VP guidance`
    $v_i \leftarrow$ VP-Selector(x_i, q_i)
    $(x_i', q_i') \leftarrow (\texttt{VP}(x_i; v_i),\ q_i \oplus \texttt{hint}(v_i))$
    for $g \leftarrow 1$ to G **do**
      $y_i'^{(g)} \sim \pi_\theta(\cdot \mid x_i', q_i'),\quad R_i'^{(g)} \leftarrow r(x_i', y_i'^{(g)})$
    end for
    $\bar{R}_i' \leftarrow \frac{1}{G}\sum_{g=1}^{G} R_i'^{(g)},\quad C_i \leftarrow \{g : R_i'^{(g)} > \bar{R}_i\}$
    if $C_i \neq \emptyset$ **then** `//* Conduct self-distillation`
      $S_i \leftarrow \texttt{TopN}(\{R_{i,\text{ans}}'^{(g)}\}_{g \in C_i})$
      $\mathcal{L}_{\text{SD}}(\theta) \leftarrow -\frac{1}{|S_i|}\sum_{j \in S_i} \frac{1}{|y_{i,j}'|}\sum_{t=1}^{|y_{i,j}'|} \log \pi_\theta(y_{i,j,t}' \mid y_{i,j,<t}', x_i', q_i')$
      $\mathcal{L}(\theta) \leftarrow \mathcal{L}_{\text{GSPO}}(\theta) + \alpha \cdot \mathbb{I}_i \cdot \mathcal{L}_{\text{SD}}(\theta)$
    end if
  else
    $\mathcal{L}(\theta) \leftarrow \mathcal{L}_{\text{GSPO}}(\theta)$
  end if
end for

facilitate spatio-temporal grounding. For example, suppressing irrelevant regions can emphasize key object areas and improve spatial reasoning. Using the prompted input (x_i', q_i'), we perform another G rollouts and obtain new reasoning y_i' and updated overall rewards $\{R_i'^{(g)}\}_{g=1}^{G}$. We expect improved reasoning trajectories and higher rewards due to the enhanced grounding signals provided by visual prompting. Detailed ablation studies on the choice of reward for candidate selection are provided in the Appendix.

Self-distillation. When visual prompting yields improved rewards, we further reinforce this improved reasoning through self-distillation. Specifically, after performing rollouts on the visual-prompted input, we identify the subset of rollouts whose overall rewards exceed the average reward of the original rollouts, i.e., $C_i = \{g \mid R_i'^{(g)} > \bar{R}_i\}$. Among these candidates, we select the top N rollouts S_i based on answering rewards. If no such candidates exist (i.e., $C_i = \emptyset$), we skip the self-distillation step for the current sample. We then apply token-level negative log-likelihood (NLL) to the selected reasoning:

$$\mathcal{L}_{\text{SD}}(\theta) = -\frac{1}{|S_i|}\sum_{j \in S_i} \frac{1}{|y_{i,j}'|}\sum_{t=1}^{|y_{i,j}'|} \log \pi_\theta(y_{i,j,t}' \mid y_{i,j,<t}', x_i', q_i'). \tag{1}$$

This objective encourages the policy to internalize high-reward reasoning trajectories generated under visual guidance. Accordingly, the final training objective is defined as:

$$\mathcal{L}_{\text{GSPO}}(\theta) + \alpha\, \mathbb{I}_i\, \mathcal{L}_{\text{SD}}(\theta), \tag{2}$$

where $\mathbb{I}_i$ is an indicator function that equals 1 if sample i is identified as a hard sample, and 0 otherwise. By repeatedly reinforcing improved trajectories, the model progressively internalizes the grounding behaviors induced by visual prompting, enabling self-evolving and more robust spatio-temporal reasoning without requiring visual prompts at inference time.

3.3 VP-Selector: Learning to Provide Visual Guidance

We now introduce VP-Selector, which is used to train the policy model ST-Reasoner within our input-adaptive RL framework. Our VP-Selector is designed to predict an input-adaptive visual prompt conditioned on the video–question pair, enabling targeted visual guidance during RL training for hard examples. Given an input (x, q), the VP-Selector selects an appropriate visual prompt v_i from a candidate pool, where each prompt provides a different form of perceptual guidance. As shown in Fig. 2 right, we first construct a training dataset using proxy reasoners [1, 19, 28] to estimate the effectiveness of candidate prompts and train the small VLM to predict the most suitable visual prompt.

Training data collection. Since defining a gold visual prompt across models is challenging, we collect (x, v_i) pairs using multiple proxy reasoners to capture general prompt effectiveness patterns. We first define a visual prompt candidate $\mathcal{V} = \{v_1, v_2, \dots\}$ by applying diverse visual prompting to key frames and objects from x. Specifically, we consider red circles [43], attention-based prompts [51], frame numbering [49], and darkening as potential visual guidance. For each candidate prompt v_k, we generate a viusal-prompted input $x'_k = \texttt{VP}(x; v_k)$ and obtain reasoning outputs $y_k = \mathcal{R}(x'_k, q)$ using multiple proxy reasoners $\mathcal{R}$ [1, 10, 28]. We compute binary answer accuracy and grounding scores (A_k, G_k), average them across proxy reasoners, and select the optimal prompt as

$$v_i = \arg\max_{v_k \in \mathcal{V}} (A_k + G_k). \tag{3}$$

The resulting (x, q, v_i) pairs are used to train the VP-Selector to predict the optimal prompt conditioned on the input video and question. Please refer to the Appendix for more details about the collected data.

Training. Given the collected dataset $\mathcal{D} = \{(x, q, v_i)\}$, we train the VP-Selector as a lightweight VLM classifier [2] with LoRA [26]. We cast prompt selection as a $|\mathcal{V}|$-way prediction problem and optimize the selector using a supervised objective. Concretely, we format the input (x, q) as an instruction to choose one method from $\mathcal{V}$ and train the model to generate the corresponding prompt label as a single-token/short-string response using token-level cross-entropy loss. The VP-Selector is frozen when incorporated in the RL framework. Additional details are provided in the Appendix.

3.4 Reward Design of ST-Reasoner

We design reward functions to train ST-Reasoner with GSPO. Specifically, we employ four reward components: (1) answer accuracy, (2) format correctness, (3) temporal grounding, and (4) object-aware spatial grounding. Among them, the object-aware spatial grounding reward is newly introduced in this work, while the remaining rewards follow prior work [38]. The overall reward is defined as

$$r(x, y) = r_{\text{acc}}(x, y) + r_{\text{fmt}}(x, y) + r_{\text{tmp}}(x, y) + r_{\text{spa}}(x, y), \tag{4}$$

and the rewards are group-normalized across rollouts to compute advantages used for GSPO updates.

Accuracy reward (r_{acc}). Following [38], we define task-specific accuracy rewards depending on the supervision type. For multiple-choice questions, the reward is binary correctness. For open-ended questions, we compute textual similarity between the predicted and ground-truth answers using ROUGE. For spatial grounding tasks, the reward is given by the visual IoU between predicted and GT bounding boxes, while for temporal grounding tasks we use temporal IoU between predicted and GT time intervals. This design provides a unified accuracy signal that adapts to heterogeneous task formats.

Format reward (r_{fmt}). We assign a binary reward that evaluates whether the generated output follows the required structured format, including the correct use of `<think>` and `<answer>` tags as well as valid `<obj>`, `<box>`, and `<t>` annotations. Outputs that satisfy the format receive a reward of 1, and 0 otherwise.

Temporal grounding reward (r_{tmp}). Following prior work [38], we set the temporal grounding reward to force the model locate correct temporal boundaries for visual evidence. Let $\mathcal{T}^{\text{gt}} \subset \mathbb{R}$ denote the set of annotated ground-truth temporal positions. Depending on the training dataset, annotations are provided either as discrete timestamps or as temporal intervals. When interval annotations are available, we denote the ground-truth interval set as $\mathcal{S}^{\text{gt}} = \{[s_\ell^{\text{gt}}, e_\ell^{\text{gt}}]\}_{\ell=1}^{L}$. From the generated reasoning y, we parse M predicted timestamps $\{t_m\}_{m=1}^{M}$. For each t_m, we first match it to the closest ground-truth temporal position:

$$\tilde{t}_m = \arg\min_{t \in \mathcal{T}^{\text{gt}}} |t_m - t|, \qquad \Delta t_m = |t_m - \tilde{t}_m|.$$

Then, we define the temporal reward as

$$r_{\text{tmp}}(x, y) = \frac{1}{M} \sum_{m=1}^{M} r_m, \tag{5}$$

where

$$r_m = \begin{cases} 1, & \text{if } \exists \ell : \ t_m \in [s_\ell^{\text{gt}}, e_\ell^{\text{gt}}], \\ \exp\Big(-\frac{\Delta t_m^2}{2\sigma^2}\Big), & \text{otherwise.} \end{cases} \tag{6}$$

Object-aware spatial grounding reward (r_{spa}). Here, we further design an object-aware spatial grounding reward to better support grounded reasoning. Prior work [38] computes spatial rewards by selecting only the predicted bounding

box with the maximum IoU, without considering object identity. However, we observe that this design encourages single-box predictions and object-agnostic hallucinations, as it does not enforce consistency between the predicted objects and the grounded region. As shown in the bottom of Fig. 2, to promote accurate object identification and multi-object grounding, we introduce a spatial reward that jointly considers soft object identity matching and all predicted box IoUs under temporal gating.

From the generated reasoning y, we parse predicted objects and bounding boxes to construct spatio-temporal tuples $\{(o_m, t_m, b_m)\}_{m=1}^{M}$, where o_m denotes the predicted object name and b_m denotes the predicted bounding box. We denote the ground-truth object set and their bounding boxes at the matched keyframe $\tilde{t}_m$ as $\mathcal{O}^{\mathrm{gt}}(\tilde{t}_m)$ and $\mathcal{B}^{\mathrm{gt}}(\tilde{t}_m)$, respectively. Using soft identity matching $\mathbb{I}_{\mathrm{soft}}(o_m, o)$, we define the identity-consistent object set as

$$\mathcal{O}_m^{\mathrm{match}} = \left\{ o \in \mathcal{O}^{\mathrm{gt}}(\tilde{t}_m) \;\middle|\; \mathbb{I}_{\mathrm{soft}}(o_m, o) = 1 \right\}. \tag{7}$$

Here $\mathbb{I}_{\mathrm{soft}}(o_m, o) = 1$ if the predicted object name o_m and the ground-truth name o satisfy one of the following conditions: (i) exact string match, (ii) substring inclusion (i.e., $o_m \subset o$ or $o \subset o_m$); otherwise $\mathbb{I}_{\mathrm{soft}}(o_m, o) = 0$.

Moreover, since spatial grounding is meaningful only when temporal localization is sufficiently accurate, we apply **temporal gating** based on the deviation Δt_m defined above. We retain only predictions that satisfy temporal alignment and identity consistency:

$$\mathcal{I} = \left\{ m \;\middle|\; \Delta t_m \le \tau \;\wedge\; |\mathcal{O}_m^{\mathrm{match}}| > 0 \right\},$$

where τ is a predefined temporal tolerance threshold. The spatial reward is then computed by averaging IoU scores over all valid predictions:

$$r_{\mathrm{spa}}(x, y) = \frac{1}{|\mathcal{I}|} \sum_{m \in \mathcal{I}} \frac{1}{|\mathcal{O}_m^{\mathrm{match}}|} \sum_{o \in \mathcal{O}_m^{\mathrm{match}}} \max_{b^{\mathrm{gt}} \in \mathcal{B}^{\mathrm{gt}}(\tilde{t}_m)} \mathrm{IoU}(b_m, b^{\mathrm{gt}}). \tag{8}$$

4 Experiments

4.1 Setup

Implementation Details. For ST-Reasoner training, we follow a two-stage pipeline consisting of SFT-based cold-start initialization followed by RL, consistent with prior work [14, 38, 45]. We first train the VP-Selector and keep it frozen during the subsequent ST-Reasoner training stage. The VP-Selector is trained using the training dataset of TVQA+ [30] and VideoEspresso [23]. For ST-Reasoner training, we adopt the same SFT and RL datasets as [38] to ensure a fair comparison. For self-distillation, we select the top-2 samples among candidates with improved rewards and set the self-distillation weight to $\alpha = 0.1$. Additional implementation details, baseline details, and threshold ablations are provided in the Appendix.

Table 2: Performance on the V-STAR benchmark. VisionCoach achieves strong spatio-temporal reasoning across overall dimensions.

Model	What	When (Temporal IoU)		Where (Spatial IoU)		Overall	
	Acc	Chain1	Chain2	Chain1	Chain2	mAM	mLGM
Proprietary Models							
Gemini-2-Flash [19]	53.0	24.5	23.8	4.6	2.2	26.9	35.6
GPT-4o [28]	60.8	16.7	12.8	6.5	3.0	26.8	38.2
Tool-calling Methods							
VideoZoomer [13]	47.7	11.1	3.7	2.0	0.5	18.8	24.6
Conan [39]	54.5	14.4	14.9	4.4	0.6	23.9	32.4
Video-o3 [54]	29.7	5.5	23.1	0.5	0.0	13.3	15.4
Tool-free Methods							
TRACE [22]	17.6	19.1	17.1	0.0	0.0	12.0	13.3
Oryx-1.5-7B [37]	20.5	13.5	14.8	10.1	3.5	15.1	13.8
VideoChat2 [34]	36.2	13.7	12.5	2.5	1.0	17.0	20.3
Qwen2.5-VL-7B* [2]	33.5	15.4	13.8	17.0	2.5	19.3	22.4
InternVL-2.5-8B [5]	44.2	8.7	7.8	0.7	0.1	17.6	24.9
Video-LLaMA3 [55]	41.9	23.0	23.1	0.9	0.2	21.7	27.0
LLaVA-Video [60]	49.5	10.5	12.2	1.9	1.3	20.8	27.3
VideoChat-R1.5 [50]	49.7	14.4	5.0	14.8	1.5	22.5	29.3
Open-o3-video* [38]	60.2	25.0	24.5	24.8	**5.9**	33.4	46.0
VisionCoach (Ours)	**61.1**	**25.7**	**25.4**	**27.2**	5.3	**34.3**	**47.5**
Δ vs. Qwen2.5-VL-7B	+27.6	+10.3	+11.6	+10.2	+2.8	+15.0	+25.1

Benchmark and Metrics. We evaluate spatio-temporal reasoning on V-STAR [7]. For zero-shot general video understanding, we report results on WorldSense [25], VideoMMMU [27], PerceptionTest [40], and VideoMME [15]. For zero-shot temporal grounding, we use Charades-STA [17]. For zero-shot spatio-temporal grounding, we use HCSTVG [44]. More details regarding metrics and benchmarks are in the Appendix.

4.2 Main Results

Results on V-STAR. We first evaluate the spatio-temporal reasoning capability of VisionCoach on V-STAR. As shown in Tab. 2, VisionCoach consistently outperforms both proprietary and open-source baselines across all evaluation dimensions. Compared to Qwen2.5-VL-7B, VisionCoach improves VQA accuracy by +**27.6** (33.5→61.1) and overall performance by +**15.0** mAM and +**25.1** mLGM. These results demonstrate that VisionCoach significantly strengthens grounded spatio-temporal reasoning, yielding consistent improvements in answer correctness, temporal localization, and spatial grounding, thereby validating the effectiveness of grounding-aware RL with self-distillation.

Results on general video understanding benchmarks. We compare the general video understanding capability of VisionCoach with GPT-4o, tool-calling models, and tool-free baselines. As shown in Tab. 3, VisionCoach consistently outperforms text-centric reasoning models (VideoR1 and VideoRFT) as well as tool-calling approaches, while remaining fully tool-free. Notably, VisionCoach shows substantial gains on perception-oriented categories, including WorldSense (Recognition) and VideoMMMU (Perception).

Table 3: Performance across different general video understanding benchmarks. * indicates our implementation performance.

Model	VideoMME		WorldSense		VideoMMMU		PerceptionTest
	Overall	Long	**Overall**	Recognition	**Overall**	Perception	**Overall**
Proprietary Models							
GPT-4o	71.9	-	42.6	-	61.2	66.0	-
Tool-calling Methods							
EgoR1	-	64.9	-	-	-	-	-
EgoR1 (w/o RAG)	58.7	50.8	42.0	40.3	34.4	37.6	65.8
LongVT-RL-7B	66.1	-	22.1	22.5	42.6	50.0	55.4
Tool-free Methods							
Qwen2.5-VL-7B	62.4	50.8	36.1	33.7	51.2	64.7	66.2
VideoRFT-7B	59.8	50.7	38.2	36.6	51.1	66.0	-
VideoR1-7B	61.4	50.6	35.5	32.8	52.4	65.3	-
Open-o3-video-7B	63.1*	53.1*	37.5	36.8	52.3	68.0	67.5
VisionCoach-7B (Ours)	**63.3**	**53.2**	**43.8**	**41.8**	**54.4**	**70.3**	**68.7**

These results indicate that improved spatio-temporal grounding not only enhances localization quality but also translates into stronger video answering.

Table 4: Zero-shot temporal grounding performance on Charades-STA.

	R@0.3	R@0.5	R@0.7	mIoU
General Video LLMs				
VideoChat	9.0	3.3	1.3	6.5
VideoLLaMA	10.4	3.8	0.9	7.1
Video-ChatGPT	20.0	7.7	1.7	13.7
Valley	28.4	1.8	0.3	21.4
Temporal Grounding Video LLM				
GroundingGPT	-	29.6	11.9	-
Momentor	42.6	26.6	11.6	28.5
TimeChat	-	32.2	13.4	-
HawkEye	50.6	31.4	14.5	33.7
VTG-LLM	51.2	33.8	15.7	34.4
VTimeLLM	51.0	27.5	11.4	31.2
Grounded Reasoning Models				
Open-o3-video	62.6	45.6	24.5	42.5
VisionCoach (Ours)	**63.2**	**45.8**	**24.7**	**42.7**

Results on the temporal grounding benchmark. We evaluate VisionCoach on a temporal grounding benchmark and compare it against existing VideoLLMs specialized in temporal grounding. As shown in Tab. 4, VisionCoach consistently outperforms all competing methods. These results indicate that the proposed visual prompt guidance effectively facilitates the learning of accurate temporal grounding during training.

Results on the spatio-temporal grounding benchmark. To further evaluate VisionCoach, we conduct experiments on HCSTVG [44], a benchmark requiring precise temporal localization and spatial grounding. As shown in Tab. 5, VisionCoach consistently outperforms existing tool-calling approaches and achieves substantial gains over Open-o3-video across all evaluation metrics.

Notably, VisionCoach remains competitive with GPT-5.4 despite being built on a significantly smaller open-source backbone. These results suggest that the proposed visual coaching framework is beneficial for fine-grained spatio-temporal reasoning, where accurate localization of relevant objects and events is crucial.

Table 5: Zero-shot spatio-temporal grounding performance on HCSTVG.

Models	m_vIoU	vIoU@0.3	vIoU@0.5
Proprietary Models			
GPT-5.4	20.2	25.9	5.1
Tool-calling Methods			
VideoChat-R1.5	11.3	11.9	2.3
VideoZoomer	7.2	5.35	0.6
Conan	6.9	5.0	0.7
Video-o3	3.8	3.6	0.3
Tool-free Methods			
Open-o3-video	17.1	22.6	5.2
VisionCoach (Ours)	**19.3** (+12.9%)	**25.4** (+12.4%)	**6.1** (+17.3%)

Table 6: Ablation of VisionCoach on V-STAR. r_{spa}: object-aware spatial grounding reward. $\mathcal{L}_{\text{SD}}$: self-distillation. VP-F: fixed visual prompting (darken). VP-S: VP-Selector.

r_{spa}	$\mathcal{L}_{\text{SD}}$	VP-F	VP-S	Acc	mAM	mLGM
✓	–	–	–	59.4	31.1	42.7
✓	✓	–	–	59.6	30.4	41.6
✓	✓	darken	–	58.3	29.7	40.6
✓	✓	-	✓	**60.7**	**31.3**	**43.1**

Table 7: VP-Selector as plug-and-play module. We evaluate the zero-shot performance of combining VP-Selector with different reasoners.

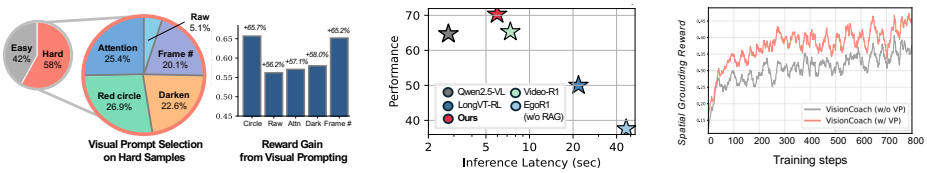
Models	TVQA+ (In-domain)	PerceptionTest (Out-domain)
Qwen2.5-VL-7B	54.5	52.2
+VP-Selector	**56.2**	**56.7**
GPT-4o	71.8	61.5
+VP-Selector	**75.3**	**62.4**
Gemini-2.5-Flash	72.4	47.4
+VP-Selector	**76.3**	**50.4**

4.3 Ablation Study

Ablation of VisionCoach: input-adaptive visual prompting is crucial for grounded reasoning. As shown in Tab. 6, we progressively enable four components of VisionCoach: (i) spatial grounding reward (r_{spa}), (ii) self-distillation loss ($\mathcal{L}_{\text{SD}}$), (iii) fixed visual prompting (VP-F), and (iv) VP-Selector (VP-S). While self-distillation slightly improves VQA accuracy, it does not consistently improve grounding performance. Fixed visual prompting further degrades grounding due to indiscriminate perturbations. In contrast, combining input-adaptive prompting with self-distillation achieves the best performance across all metrics, demonstrating the importance of selectively applying visual guidance. Furthermore, VP-guided training yields higher rewards and faster convergence (Fig. 4c), suggesting that visual prompting provides effective grounding signals during RL training. Additional ablation studies are provided in the Appendix.

Statistics of adaptive visual prompting: adaptive visual prompting effectively targets difficult samples and provides strong learning signals. To better understand the behavior of VisionCoach's adaptive visual prompting mechanism, we analyze the distribution of designated sample difficulty (Easy vs. Hard), the visual prompts selected for hard samples by VP-Selector, and the resulting reward gain from each prompt. We compute the relative reward gain as $(R'_i - \bar{R}_i)/\bar{R}_i \times 100$. As shown in Fig. 4 (a), 58% of the samples are identified as hard. For these hard samples, VP-Selector dynamically chooses among multiple visual prompting strategies. Importantly, applying these prompts consistently leads to reward improvements, with gains ranging from +56% to +66% across different prompt types. These results indicate that visual prompting provides effective guidance for challenging instances and that the selector successfully identifies prompts that yield meaningful training signals.

VP-Selector as a plug-and-play module: adaptive visual prompting improves video reasoning across model-agnostic backbones. As shown in Tab. 7, we evaluate VP-Selector as a standalone module by integrating it with various MLLM backbones to augment video inputs prior to reasoning. We conduct experiments on the TVQA+ validation set (in-domain) and on a uniformly sampled subset from each category of the PerceptionTest validation set (out-of-domain). Incorporating VP-Selector consistently improves performance



(a) Statistics of input-adaptive visual prompting

(b) Comparison of inference latency

(c) VP improves spatial grounding reward

Fig. 4: Additional analysis of VisionCoach. We provide analysis including (a) statistics of visual prompting, (b) inference latency, and (c) the effect of visual prompting on spatial grounding reward.

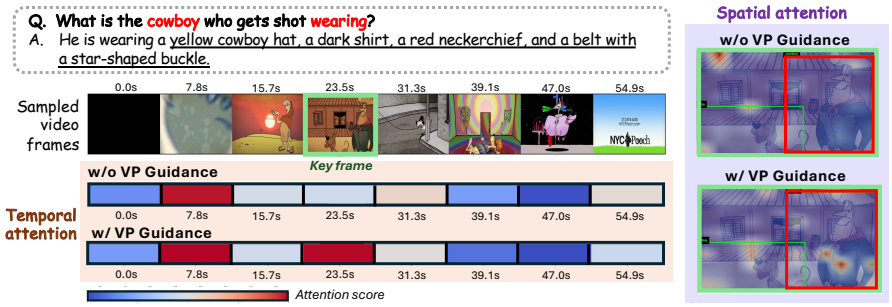


Fig. 5: Spatio-temporal attention map. Visual prompting (VP) improves grounding in both temporal and spatial dimensions. The green box indicates the key frame, while the red box highlights the corresponding spatial region. Temporally, VP increases attention on the correct key frame containing the cowboy. Spatially, VP concentrates attention on the region corresponding to the queried visual attributes (e.g., the cowboy wearing specific clothing), while suppressing irrelevant regions.

across different backbone models on both benchmarks. These results demonstrate that VP-Selector functions as a plug-and-play module for enhancing grounded visual reasoning.

4.4 More Analysis

Inference latency: VisionCoach achieves higher performance with lower inference latency. To evaluate per-sample inference latency, we measure runtime on a single NVIDIA RTX 6000 (40GB) GPU and compare text-centric reasoning, tool-calling reasoning, and VisionCoach. As shown in Fig. 4 (b), VisionCoach consistently outperforms both text-centric (Qwen2.5-VL, Video-R1) and tool-calling (EgoR1, LongVT-RL) baselines while operating at substantially lower inference latency than external tool-based approaches.

Spatio-temporal attention map: VP guidance encourages more focused temporal and spatial attention on visual evidence. To examine how visual prompting facilitates spatio-temporal grounding during training, we analyze the

attention maps from the last layer of VisionCoach. Specifically, we compute the average attention across all heads and measure the attention scores between the question tokens and visual tokens. Temporal attention is aggregated at the frame level, while spatial attention is aggregated over patch-level regions within each frame. As illustrated in Fig. 5, without VP guidance, both spatial and temporal attention are broadly dispersed across irrelevant frames and background regions, resulting in weak localization of the question-relevant event. In contrast, with VP guidance, the model concentrates attention on the key frame where the cowboy is shot and focuses on semantically relevant regions, such as the character and surrounding objects. This comparison indicates that VP guidance promotes more accurate temporal localization and spatial grounding, enabling better alignment between reasoning and visual evidence. Additional analyses, including qualitative examples, reward ablations, and limitations are provided in the Appendix.

5 Conclusion

In this work, we introduced VisionCoach, an instance-adaptive RL framework for grounded video reasoning. VisionCoach applies visual prompting during RL to explicitly expose question-relevant evidence during learning using VP-Selector and ST-Reasoner. Experiments across multiple challenging benchmarks across tasks show consistent gains in both reasoning accuracy and grounding quality, while preserving a single, efficient inference pathway.

Acknowledgment

This work was supported by ONR Grant N00014-23-1-2356, ARO Award W911-NF2110220, DARPA ECOLE Program No. HR00112390060, and NSF-AI Engage Institute DRL2112635. The views contained in this article are those of the authors and not of the funding agency.

References

1. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., et al.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025)
2. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., et al.: Qwen2. 5-vl technical report. corr, abs/2502.13923, 2025. doi: 10.48550. arXiv preprint 2502.13923 (2025)
3. Cai, M., Liu, H., Mustikovela, S.K., Meyer, G.P., Chai, Y., Park, D., Lee, Y.J.: Vip-llava: Making large multimodal models understand arbitrary visual prompts pp. 12914–12923 (2024)
4. Chen, Y., Huang, W., Shi, B., Hu, Q., Ye, H., Zhu, L., Liu, Z., Molchanov, P., Kautz, J., Qi, X., et al.: Scaling rl to long videos. Advances in Neural Information Processing Systems **38**, 172842–172870 (2026)

5. Chen, Z., Wu, J., Wang, W., Su, W., Chen, G., Xing, S., Zhong, M., Zhang, Q., Zhu, X., Lu, L., et al.: Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 24185–24198 (2024)
6. Cheng, J., Ge, Y., Wang, T., Ge, Y., Liao, J., Shan, Y.: Video-holmes: Can mllm think like holmes for complex video reasoning? arXiv preprint arXiv:2505.21374 (2025)
7. Cheng, Z., Hu, J., Liu, Z., Si, C., Li, W., Gong, S.: V-star: Benchmarking video-llms on video spatio-temporal reasoning. arXiv preprint arXiv:2503.11495 (2025)
8. Cheng, Z., Li, D., Hu, J., Liu, Z., Li, W., Gong, S.: Graphthinker: Reinforcing video reasoning with event graph thinking. arXiv preprint arXiv:2602.17555 (2026)
9. Choudhury, R., Niinuma, K., Kitani, K.M., Jeni, L.A.: Video question answering with procedural programs. In: European Conference on Computer Vision. pp. 315–332. Springer (2024)
10. Comanici, G., Bieber, E., Schaekermann, M., Pasupat, I., Sachdeva, N., Dhillon, I., Blistein, M., Ram, O., Zhang, D., Rosen, E., 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)
11. Deng, A., Chen, T., Yu, S., Yang, T., Spencer, L., Tian, Y., Mian, A.S., Bansal, M., Chen, C.: Motion-grounded video reasoning: Understanding and perceiving motion at pixel level. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 8625–8636 (2025)
12. Deng, A., Yang, T., Yu, S., Spencer, L., Bansal, M., Chen, C., Yeung-Levy, S., Wang, X.: Scivideobench: Benchmarking scientific video reasoning in large multimodal models. arXiv preprint arXiv:2510.08559 (2025)
13. Ding, Y., Zhang, Y., Lai, X., Chu, R., Yang, Y.: Videozoomer: Reinforcement-learned temporal focusing for long video reasoning. arXiv preprint arXiv:2512.22315 (2025)
14. Feng, K., Gong, K., Li, B., Guo, Z., Wang, Y., Peng, T., Wu, J., Zhang, X., Wang, B., Yue, X.: Video-r1: Reinforcing video reasoning in mllms. arXiv preprint arXiv:2503.21776 (2025)
15. Fu, C., Dai, Y., Luo, Y., Li, L., Ren, S., Zhang, R., Wang, Z., Zhou, C., Shen, Y., Zhang, M., et al.: Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 24108–24118 (2025)
16. Furlanello, T., Lipton, Z., Tschannen, M., Itti, L., Anandkumar, A.: Born again neural networks. In: International conference on machine learning. pp. 1607–1616. PMLR (2018)
17. Gao, J., Sun, C., Yang, Z., Nevatia, R.: Tall: Temporal activity localization via language query. In: Proceedings of the IEEE international conference on computer vision. pp. 5267–5275 (2017)
18. Ge, H., Wang, Y., Chang, K.W., Wu, H., Cai, Y.: Framemind: Frame-interleaved video reasoning via reinforcement learning. arXiv preprint arXiv:2509.24008 (2025)
19. Google DeepMind: Gemini 2 technical report (2025), `https://deepmind.google/technologies/gemini/`
20. Gu, X., Zhang, H., Fan, Q., Niu, J., Zhang, Z., Zhang, L., Chen, G., Chen, F., Wen, L., Zhu, S.: Thinking with bounding boxes: Enhancing spatio-temporal video grounding via reinforcement fine-tuning. arXiv preprint arXiv:2511.21375 (2025)
21. Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X., et al.: Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. arXiv preprint arXiv:2501.12948 (2025)

22. Guo, Y., Liu, J., Li, M., Liu, Q., Chen, X., Tang, X.: Trace: Temporal grounding video llm via causal event modeling. arXiv preprint arXiv:2410.05643 (2024)
23. Han, S., Huang, W., Shi, H., Zhuo, L., Su, X., Zhang, S., Zhou, X., Qi, X., Liao, Y., Liu, S.: Videoespresso: 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. pp. 26181–26191 (2025)
24. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network (2015)
25. Hong, J., Yan, S., Cai, J., Jiang, X., Hu, Y., Xie, W.: Worldsense: Evaluating real-world omnimodal understanding for multimodal llms. arXiv preprint arXiv:2502.04326 (2025)
26. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: Lora: Low-rank adaptation of large language models. Iclr **1**(2), 3 (2022)
27. Hu, K., Wu, P., Pu, F., Xiao, W., Zhang, Y., Yue, X., Li, B., Liu, Z.: Video-mmmu: Evaluating knowledge acquisition from multi-discipline professional videos. arXiv preprint arXiv:2501.13826 (2025)
28. Hurst, A., Lerer, A., Goucher, A.P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., et al.: Gpt-4o system card. arXiv preprint arXiv:2410.21276 (2024)
29. Lee, D., Yoon, J., Cho, J., Bansal, M.: Video-skill-cot: Skill-based chain-of-thoughts for domain-adaptive video reasoning. arXiv preprint arXiv:2506.03525 (2025)
30. Lei, J., Yu, L., Berg, T., Bansal, M.: Tvqa+: Spatio-temporal grounding for video question answering. In: Proceedings of the 58th annual meeting of the association for computational linguistics. pp. 8211–8225 (2020)
31. Li, B., Zhang, Y., Guo, D., Zhang, R., Li, F., Zhang, H., Zhang, K., Zhang, P., Li, Y., Liu, Z., et al.: Llava-onevision: Easy visual task transfer. arXiv preprint arXiv:2408.03326 (2024)
32. Li, F., Jiang, Q., Zhang, H., Ren, T., Liu, S., Zou, X., Xu, H., Li, H., Yang, J., Li, C., et al.: Visual in-context prompting pp. 12861–12871 (2024)
33. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: International conference on machine learning. pp. 19730–19742. PMLR (2023)
34. Li, K., Wang, Y., He, Y., Li, Y., Wang, Y., Liu, Y., Wang, Z., Xu, J., Chen, G., Luo, P., et al.: Mvbench: A comprehensive multi-modal video understanding benchmark. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22195–22206 (2024)
35. Li, X., Yan, Z., Meng, D., Dong, L., Zeng, X., He, Y., Wang, Y., Qiao, Y., Wang, Y., Wang, L.: Videochat-r1: Enhancing spatio-temporal perception via reinforcement fine-tuning. arXiv preprint arXiv:2504.06958 (2025)
36. Liu, A., Feng, B., Xue, B., Wang, B., Wu, B., Lu, C., Zhao, C., Deng, C., Zhang, C., Ruan, C., et al.: Deepseek-v3 technical report. arXiv preprint arXiv:2412.19437 (2024)
37. Liu, Z., Dong, Y., Liu, Z., Hu, W., Lu, J., Rao, Y.: Oryx mllm: On-demand spatial-temporal understanding at arbitrary resolution. arXiv preprint arXiv:2409.12961 (2024)
38. Meng, J., Li, X., Wang, H., Tan, Y., Zhang, T., Kong, L., Tong, Y., Wang, A., Teng, Z., Wang, Y., et al.: Open-o3 video: Grounded video reasoning with explicit spatio-temporal evidence. arXiv preprint arXiv:2510.20579 (2025)
39. Ouyang, K., Liu, Y., Yao, L., Cai, Y., Zhou, H., Meng, F., Zhou, J., Sun, X.: Conan: Progressive learning to reason like a detective over multi-scale visual evidence.

In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 41089–41099 (2026)
40. Patraucean, V., Smaira, L., Gupta, A., Recasens, A., Markeeva, L., Banarse, D., Koppula, S., Malinowski, M., Yang, Y., Doersch, C., et al.: Perception test: A diagnostic benchmark for multimodal video models. Advances in Neural Information Processing Systems **36**, 42748–42761 (2023)
41. Rasheed, H., Zumri, M., Maaz, M., Yang, M.H., Khan, F.S., Khan, S.: Videocom: Interactive video reasoning via chain of manipulations. arXiv preprint arXiv:2511.23477 (2025)
42. Rusu, A.A., Colmenarejo, S.G., Gulcehre, C., Desjardins, G., Kirkpatrick, J., Pascanu, R., Mnih, V., Kavukcuoglu, K., Hadsell, R.: Policy distillation. arXiv preprint arXiv:1511.06295 (2015)
43. Shtedritski, A., Rupprecht, C., Vedaldi, A.: What does clip know about a red circle? visual prompt engineering for vlms pp. 11987–11997 (2023)
44. Tang, Z., Liao, Y., Liu, S., Li, G., Jin, X., Jiang, H., Yu, Q., Xu, D.: Human-centric spatio-temporal video grounding with visual transformers. IEEE Transactions on Circuits and Systems for Video Technology **32**(12), 8238–8249 (2021)
45. Wang, Q., Yu, Y., Yuan, Y., Mao, R., Zhou, T.: Videorft: Incentivizing video reasoning capability in mllms via reinforced fine-tuning. arXiv preprint arXiv:2505.12434 (2025)
46. Wang, Z., Yoon, J., Yu, S., Islam, M.M., Bertasius, G., Bansal, M.: Video-rts: Rethinking reinforcement learning and test-time scaling for efficient and enhanced video reasoning. In: Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing. pp. 28114–28128 (2025)
47. Wu, B., Yu, S., Chen, Z., Tenenbaum, J.B., Gan, C.: Star: A benchmark for situated reasoning in real-world videos. arXiv preprint arXiv:2405.09711 (2024)
48. Wu, J., Zhang, Z., Xia, Y., Li, X., Xia, Z., Chang, A., Yu, T., Kim, S., Rossi, R.A., Zhang, R., et al.: Visual prompting in multimodal large language models: A survey. arXiv preprint arXiv:2409.15310 (2024)
49. Wu, Y., Hu, X., Sun, Y., Zhou, Y., Zhu, W., Rao, F., Schiele, B., Yang, X.: Number it: Temporal grounding videos like flipping manga. arXiv preprint arXiv:2411.10332 (2024)
50. Yan, Z., Li, X., He, Y., Yue, Z., Zeng, X., Wang, Y., Qiao, Y., Wang, L., Wang, Y.: Videochat-r1.5: Visual test-time scaling to reinforce multimodal reasoning by iterative perception. arXiv preprint arXiv:2509.21100 (2025)
51. Yu, R., Yu, W., Wang, X.: Attention prompting on image for large vision-language models. In: European Conference on Computer Vision. pp. 251–268. Springer (2024)
52. Yu, S., Zhang, Y., Wang, Z., Yoon, J., Yao, H., Ding, M., Bansal, M.: When and how much to imagine: Adaptive test-time scaling with world models for visual spatial reasoning. arXiv preprint arXiv:2602.08236 (2026)
53. Zeng, X., Zhang, Z., Zhu, Y., Li, X., Wang, Z., Ma, C., Zhang, Q., Huang, Z., Ouyang, K., Jiang, T., et al.: Video-o3: Native interleaved clue seeking for long video multi-hop reasoning. arXiv preprint arXiv:2601.23224 (2026)
54. Zeng, X., Zhang, Z., Zhu, Y., Li, X., Wang, Z., Ma, C., Zhang, Q., Huang, Z., Ouyang, K., Jiang, T., et al.: Video-o3: Native interleaved clue seeking for long video multi-hop reasoning. arXiv preprint arXiv:2601.23224 (2026)
55. Zhang, B., Li, K., Cheng, Z., Hu, Z., Yuan, Y., Chen, G., Leng, S., Jiang, Y., Zhang, H., Li, X., et al.: Videollama 3: Frontier multimodal foundation models for image and video understanding. arXiv preprint arXiv:2501.13106 (2025)

56. Zhang, H., Gu, X., Li, J., Ma, C., Bai, S., Zhang, C., Zhang, B., Zhou, Z., He, D., Tang, Y.: Thinking with videos: Multimodal tool-augmented reinforcement learning for long video reasoning. arXiv preprint arXiv:2508.04416 (2025)
57. Zhang, J., Guo, Y., Potamias, R.A., Deng, J., Xu, H., Ma, C.: Vtimecot: Thinking by drawing for video temporal grounding and reasoning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 24203–24213 (2025)
58. Zhang, Y., Dong, Y., Zhang, S., Min, T., Su, H., Zhu, J.: Exploring the transferability of visual prompting for multimodal large language models pp. 26562–26572 (2024)
59. Zhang, Y., Fan, C.K., Huang, T., Lu, M., Yu, S., Pan, J., Cheng, K., She, Q., Zhang, S.: Autov: Learning to retrieve visual prompt for large vision-language models. arXiv preprint arXiv:2506.16112 (2025)
60. Zhang, Y., Wu, J., Li, W., Li, B., Ma, Z., Liu, Z., Li, C.: Llava-video: Video instruction tuning with synthetic data. arXiv preprint arXiv:2410.02713 (2024)
61. Zheng, C., Liu, S., Li, M., Chen, X.H., Yu, B., Gao, C., Dang, K., Liu, Y., Men, R., Yang, A., et al.: Group sequence policy optimization. arXiv preprint arXiv:2507.18071 (2025)