

Are Video Reasoning Models Ready to Go Outside?

Yangfan He¹, Changgyu Boo², and Jaehong Yoon^{1*}

¹NTU Singapore ²Korea University

yhe032@e.ntu.edu.sg, jaehong.yoon@ntu.edu.sg

Project Page: <https://robust-video-reason.github.io/>

Abstract. In real-world deployment, vision-language models often encounter disturbances such as weather, occlusion, and camera motion. Under such conditions, their understanding and reasoning degrade substantially, revealing a gap between clean, controlled (i.e., unperturbed) evaluation settings and real-world robustness. To address this limitation, we propose ROVA, a novel training framework that improves robustness by modeling a robustness-aware consistency reward under spatio-temporal corruptions. ROVA introduces a difficulty-aware online training strategy that prioritizes informative samples based on the model’s evolving capability. Specifically, it continuously re-estimates sample difficulty via self-reflective evaluation, enabling adaptive training with a robustness-aware consistency reward. We also introduce PVRBench, a new benchmark that injects real-world perturbations into embodied video datasets to assess both accuracy and reasoning quality under realistic disturbances. We evaluate ROVA and baselines on PVRBench, UrbanVideo, and VisBench, where open-source and proprietary models suffer up to 35% and 28% drops in accuracy and reasoning under realistic perturbations. ROVA effectively mitigates performance degradation, boosting relative accuracy by at least 24% and reasoning by over 9% compared with baseline models (QWen2.5/3-VL, InternVL2.5, Embodied-R). These gains transfer to clean standard benchmarks, yielding consistent improvements.

Keywords: video-language models · video reasoning · embodied AI · model generalization and robustness

1 Introduction

Vision-language models (VLMs) [6, 13, 14, 23, 33, 35, 37] have rapidly advanced video understanding and reasoning, allowing systems to interpret complex scenes and perform temporally grounded inference. These capabilities support many real-world applications, yet a key question remains: *are current VLMs robust enough to operate reliably beyond clean, controlled conditions?* In practice, these models frequently face challenging video streams, corrupted by adverse weather (e.g., rain, fog, snow), dynamic occlusions (e.g., pedestrians, vehicles, vegetation), abrupt illumination changes (e.g., glare, shadows, low light), and camera motion

* Corresponding author

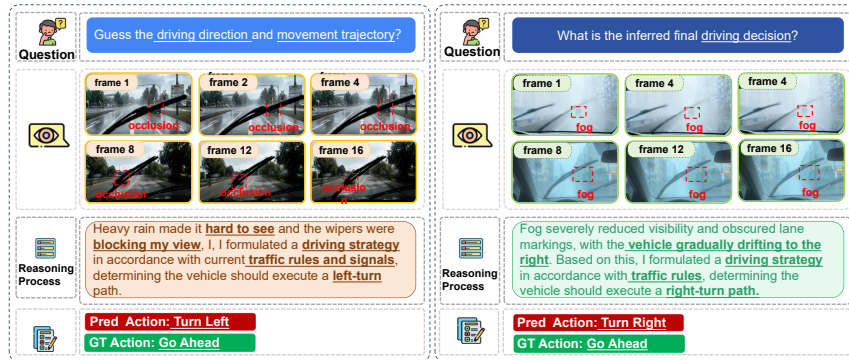


Fig. 1: Failure cases of the baseline model (Qwen2.5-VL) under two representative perturbations: (a) occlusion (left) and (b) adverse weather (right). The model incorrectly predicts “Turn Left” under occlusion and “Turn Right” under fog, despite the ground-truth being “Go Ahead” in both cases, demonstrating how realistic perturbations mislead reasoning and motivating the need for robustness-aware training.

induced by vibration or viewpoint shifts. Such perturbations are common in the real world, yet these models severely degrade perception and lead to brittle or unreliable reasoning (Fig. 1). For instance, under conditions such as video occlusion or adverse weather, baseline models may incorrectly output “Turn Left” or “Turn Right” rather than the ground-truth “Going Ahead.” This gap between benchmark assumptions and real-world conditions highlights the need for training frameworks that promote reliable generalization under realistic variability and uncertainty. A few prior studies [15, 38, 45] have explored improving the robustness of VLMs through generic data augmentation, random frame masking, zero-shot, or adversarial training. However, these methods typically treat robustness as a single objective, overlooking that different perturbations induce distinct failure modes. Consequently, they struggle to address structured, semantically meaningful corruptions common in real-world environments, since perturbation-specific failure behaviors are not explicitly modeled.

To address this challenge, we propose *RObust Video Alignment (ROVA)*, a novel training approach for robust vision reasoning under realistic visual disturbances. We first apply corruption-based augmentation to generate perturbed videos. ROVA then measures divergence in reasoning coherence and answer quality between clean and corrupted videos as a proxy for corruption-induced difficulty. Moderately difficult instances are used for training, while overly easy samples are discarded and excessively difficult ones are stored in a temporal memory buffer for later revisiting. Unlike curriculum learning, which follows a fixed, easy-to-hard schedule, this *self-reflective evaluation* estimates the difficulty and informativeness of each video-query instance based on the model’s current capability, enabling an adaptive curriculum that prioritizes informative samples while deferring overly difficult ones through memory replay. Next, we introduce a dual-branch alignment objective that enforces output consistency between paired clean and perturbed inputs. This robustness-aware consistency alignment is guided by reward modeling over reasoning and answer consistency,

Table 1: Comparison of PVRBench with existing video understanding benchmarks. *#Typ.* counts perturbation subtypes. *#Cat.* counts scene or class categories. ***Synthetic***, ***Spatial***, and ***Temporal*** indicate artificially generated, spatially grounded, and temporally consistent perturbations, respectively. PVRBench covers 27 tasks covering ***indoor***, ***outdoor***, and ***embodied AI*** scenarios. ‡: An image-level benchmark for reference.

Benchmark	Scale		Perturbation Properties					Scene Coverage			
	#Vid.	#QA	Syn.	Real	Spat.	Temp.	#Typ.	In.	Out.	Emb.	#Cat.
ImageNet-C‡ [29]	50K	50K	✓	✗	✗	✗	19	✓	✓	✗	1K
MVBench [12]	4K	4K	✗	✗	✗	✗	0	✓	✓	✗	20
Video-MME [8]	900	2.7K	✗	✗	✗	✗	0	✓	✓	✗	30
ALFRED [22]	8K	25K	✗	✗	✗	✗	0	✓	✗	✓	7
Ego4D [9]	3.7K	3.8M	✗	✗	✗	✗	0	✓	✓	✓	5
VisBench [30]	500	3K	✗	✗	✗	✗	0	✓	✗	✓	11
UrbanVideo [40]	1.5K	6K	✗	✗	✗	✗	0	✗	✓	✓	16
PVRBench (Ours)	9K	52K	✓	✗	✓	✓	12	✓	✓	✓	27

and optimized using group relative policy optimization. Specifically, we enforce output consistency between paired clean and perturbed video inputs through reward-guided optimization that evaluates both reasoning and answer consistency, trained via group relative policy optimization [20].

We further introduce *Perturbed Video Reasoning Benchmark (PVRBench)*, for evaluating the robustness of video reasoning under diverse realistic perturbations. Unlike existing benchmarks, including VisBench [30] and UrbanVideo [40], which primarily evaluate models on curated environments, PVRBench systematically injects perturbations from 12 corruption styles associated with lighting, camera motion, occlusion, and weather (Tab. 1), across 27 scene categories. Notably, all perturbations are spatially aware and temporally coherent, capturing realistic video disturbances. We observe that performant proprietary models (GPT-4o [11] / Gemini-3-Pro [24]) suffer 11–17% and 10–14% drops in accuracy and reasoning, and open-source models degrade by up to 35% and 26%, respectively, highlighting robustness gaps in VLMs under realistic conditions.

ROVA consistently outperforms proprietary and open-source models on PVRBench, UrbanVideo, and VisBench across all perturbation types in both answer accuracy and reasoning quality. Specifically, ROVA surpasses the strongest open-source baselines of comparable size, Embodied-R, by 17%, while larger variants (13B/72B) match or exceed leading proprietary models such as Gemini-3-Pro and GPT-4o. Notably, these improvements extend to clean videos, demonstrating enhanced generalizability and stronger performance on clean data. Furthermore, ROVA achieves higher reasoning quality, with improved consistency and belief scores, reflecting more stable, confident reasoning under visual corruption.

2 Related Work

Robust Training for Multimodal Models. Several works [1, 16, 18, 19, 21, 43] have explored robustness to distribution shifts and adversarial inputs through data augmentation [7], test-time adaptation [42], and transfer-based strate-

gies [3, 25]. However, these approaches primarily address generic perturbations or optimization efficiency, rather than the structured, semantically grounded disturbances encountered in real-world video settings. In video reasoning, recent methods [4, 26, 27, 46] improve efficiency via adaptive frame sampling or data filtering, but they do not explicitly model realistic corruption patterns [31, 36] that alter scene visibility and temporal coherence. As a result, robustness is treated as incidental resilience rather than being explicitly modeled during optimization. In contrast, ROVA incorporates structured and semantically grounded perturbations that reflect realistic environmental disturbances. The proposed architecture and training objectives enforce representation consistency between clean and perturbed videos, progressively strengthening disturbance-aware reasoning.

Robust Video Reasoning in Real-World Environments. Recent advances in video–language models [6, 17, 33, 35, 37] have substantially improved temporal reasoning and long-horizon embodied planning [2, 5, 32, 34, 39, 41]. However, most existing benchmarks evaluate models under nearly clean visual conditions [14], implicitly assuming stable lighting, unobstructed views, and smooth camera movement. Although robustness is sometimes measured via synthetic textual perturbations [28], such evaluations do not capture structured, semantically grounded visual disturbances encountered in real-world environments. Consequently, no standardized benchmark systematically integrates realistic disturbances into embodied video reasoning, leaving a gap between benchmarks and deployment conditions. In contrast, we introduce PVRBench that integrates semantically meaningful perturbations into temporally coherent reasoning tasks. Rather than treating corruption as incidental noise, we ask models to reliably reason about scene content, even in the presence of disturbances.

3 Training Robust Video Reasoning Models with ROVA

As illustrated in Fig. 2, ROVA, a novel training approach for robust video reasoning under real-world perturbations, comprises three stages: we first generate corruption-augmented video-query pairs via dynamic, physically plausible perturbations (Sec. 3.1). Next, a difficulty-aware curriculum performs self-reflective evaluation to selectively curate informative training samples conditioned on the model’s evolving capability (Sec. 3.2). Finally, dual-branch alignment enforces consistency between clean and perturbed videos via reasoning-aware rewards and group relative policy optimization (GRPO) (Sec. 3.3).

3.1 Learning with Structured Spatio-Temporal Corruption

We first design a structured spatio-temporal corruption pipeline that models four realistic disturbances, including *weather*, *lighting*, *occlusion*, and *camera motion*, using style-specific, cross-frame coherent masks for spatial perturbations and temporal shuffling to disrupt temporal order. Unlike generic augmentations that apply independent pixel or frame perturbations (e.g., random masking, color jittering) [29], we explicitly model perturbation styles with spatial grounding

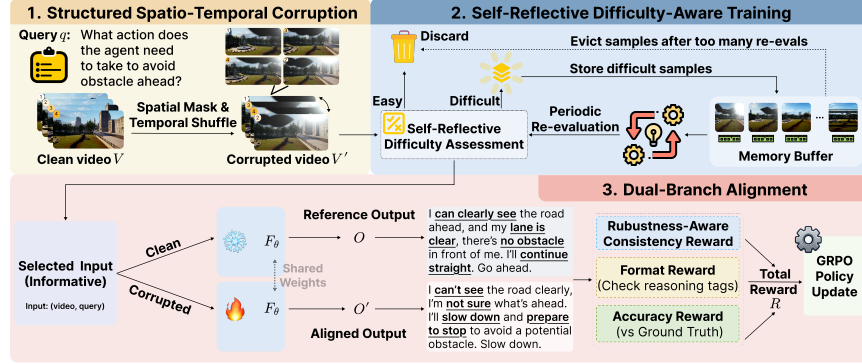


Fig. 2: Overview of ROVA: (1) structured spatio-temporal corruption that generates realistic perturbations, (2) self-reflective evaluation with difficulty-aware online training that adaptively prioritizes informative samples, and (3) dual-branch alignment reward modeling that enforces output consistency between clean and perturbed inputs.

and temporal coherence, yielding structured spatio-temporal disturbances. Each video is then paired with its corrupted counterpart in a dual-branch alignment framework to optimize output consistency. Through this design, the model learns perturbation-invariant representations for robust real-world generalization.

Let a video sequence be denoted as $V = \{f_1, f_2, \dots, f_T\}$, where $f_t \in \mathbb{R}^{H \times W \times C}$ denotes the t -th frame of height H , width W , and C channels.

Temporal Corruption. To disrupt temporal coherence, we randomly permute the frame sequence. A permutation $\pi : \{1, \dots, T\} \rightarrow \{1, \dots, T\}$ is sampled uniformly at random, and the temporally shuffled video is defined as

$$V_{\text{temp}} = \{f_{\pi(1)}, f_{\pi(2)}, \dots, f_{\pi(T)}\}, \quad (1)$$

which completely scrambles temporal order while preserving all frame content.

Spatial Corruption. Rather than coarse block-wise masking that risks removing critical cues, we apply fine-grained masks across four perturbation styles $m \in \mathcal{P} = \{\text{weather, lighting, camera, occlusion}\}$. For each frame f_t , the mask $P_t^{(m)} = B_t^{(m)} \odot C_t^{(m)}$ fuses a binary map $B_t^{(m)} \in \{0, 1\}^{H \times W}$, where 1/0 denotes corrupted/clean pixels, with layouts driven by depth awareness or stochastic sampling, and a continuous modulation map $C_t^{(m)} \in [0, 1]^{H \times W}$ encoding per-pixel effect intensity (e.g., rain strength, shadow depth, blur kernel; see more details in the appendix.) The corrupted frame is computed as $f_t^{\text{masked}} = f_t \odot P_t^{(m)}$, where $\odot$ denotes element-wise multiplication.

Spatio-Temporal Corruption. For each video, a perturbation style $m \in \mathcal{P}$ is uniformly sampled to generate the corrupted frame sequence:

$$V' = \left\{ f_{\pi(t)} \odot P_t^{(m)} \right\}_{t=1}^T, \quad (2)$$

where $P_t^{(m)}$ denotes the smooth, style-specific mask associated with style m . By jointly introducing temporal order disruption and spatially realistic, continuous

masking, our approach promotes perturbation-invariant representation learning while preserving essential visual semantics.

3.2 Self-Reflective Difficulty-Aware Training

Introducing structured visual corruptions exposes the model to a broader spectrum of reasoning difficulty than training on clean videos alone. While clean inputs typically lie within a narrow difficulty range, corrupted versions vary widely in severity, expanding the diversity of learning signals during training. Crucially, training is most effective on samples that are neither too easy nor excessively difficult [27] under the model’s current capacity, as these instances provide the most informative learning signals and support stable optimization. Rather than uniformly sampling across the expanded difficulty range, we therefore prioritize appropriately challenging examples through a self-reflective, difficulty-aware strategy that implicitly forms an online curriculum. By continuously focusing on corrupted samples that provide meaningful learning signals, the model enables to promote robust and reliable reasoning under realistic visual disturbances.

To this end, we propose a self-reflective, difficulty-aware training pipeline that implicitly builds an adaptive curriculum in an online manner. Formally, let F_θ denote a learnable VLM parameterized by θ . We assume that training video–text pairs arrive sequentially, and let θ_i denote the model parameters at training iteration i . At each iteration, ROVA performs two internal steps:

- 1) *self-reflective evaluation*, where F_{θ_i} estimates the usefulness of incoming samples for training under its current state; and
- 2) *difficulty-aware selective training*, where model updates are performed using only a subset of samples selected according to the proposed policy.

Self-Reflective Evaluation. At iteration i , the current model F evaluates each masked video V'_i and produces a difficulty label $d \in \{easy, difficult, informative\}$ and a confidence score $c \in [0, 1]$, defined as,

$$d, c = F_{\theta_i}(q_i, V'_i, S_e), \quad (3)$$

where q_i denotes the input query and S_e denotes the evaluation prompt (See details in the appendix). Specifically, d is obtained by prompting F_{θ_i} with S_e to compare its responses on clean and corrupted inputs: if the model answers correctly and consistently, the sample is labeled $d = easy$; if responses diverge substantially or are incorrect, it is labeled $d = difficult$; otherwise, the sample is labeled $d = informative$, indicating moderate uncertainty that is most beneficial for training. The confidence score c is derived from the model’s output token probabilities. Unlike traditional curriculum learning with a fixed schedule, our prompt-based sample-level evaluation dynamically estimates the model’s current capability and prioritizes informative samples to stabilize the effective training distribution. Based on d and c , we design the following data selection policy:

- (i) high-confidence easy samples ($d = easy, c > \tau$, where τ is a threshold) are considered as sufficiently learned and filtered out, enabling the model to prioritize disturbance-sensitive samples that provide strong learning signals.

- (ii) difficult samples ($d = \textit{difficult}$) are stored in a temporal memory buffer $\mathcal{M}$ for deferred training and periodically re-evaluated. While potentially informative, they may yield weak or unstable learning signals under the current model state, and are revisited once the model has sufficiently improved.
- (iii) informative samples ($d = \textit{informative}$) as well as low-confidence easy samples ($d = \textit{easy}, c \leq \tau$) are treated as high-information instances and prioritized for immediate training.

Difficulty Re-evaluation and Deferred Training with Memory. As the model improves over time, samples that were previously too difficult to learn from may later provide meaningful training signals. To leverage this evolving capability, we introduce a memory-based deferred training mechanism that periodically re-evaluates difficult instances. Formally, when newly arriving data are evaluated as *difficult*, it is stored in a temporal memory buffer $\mathcal{M}$ as:

$$\mathcal{M} \leftarrow \mathcal{M} \cup \{(q, \tilde{V}, k = 0)\}, \quad (4)$$

where $\tilde{V}$ encodes the mask metadata, including perturbation style, parameters, and spatial-temporal regions. This design allows the corrupted video V' to be regenerated on demand during re-evaluation, avoiding the need to store full video data. During training, instances in $\mathcal{M}$ are periodically re-evaluated under the updated model. The counter k tracks the number of re-assessments performed for each sample. For each entry $(q_n, \tilde{V}_n, k_n) \in \mathcal{M}$, the current model F periodically re-assesses its difficulty using the current parameter θ_i :

$$d', c' = F(q_n, \tilde{V}_n, S_e; \theta_i), \quad k_n \leftarrow k_n + 1. \quad (5)$$

Here, d' and c' denote the updated difficulty level and confidence score, respectively. Entries reclassified as *informative* are immediately used for training, whereas those labeled *easy* are removed from the memory buffer. Entries that remain difficult are retained in $\mathcal{M}$ with their re-evaluation counter incremented.

The confidence score c' serves as an auxiliary diagnostic signal for self-monitoring and stability analysis, but is not used directly for memory retention decisions to avoid sensitivity to noisy confidence estimates. As training progresses, samples that were previously difficult may transition to informative or easy categories, allowing the curriculum to adapt to the model’s evolving capability. However, repeated re-evaluation can lead to unbounded memory growth, particularly when samples remain persistently difficult or heavily corrupted, yielding little effective learning signal. To prevent this, we impose a maximum re-evaluation threshold and evict entries exceeding it:

$$\mathcal{M} \leftarrow \mathcal{M} \setminus \{(q, \tilde{V}, k) \mid k > K_{\max}\}. \quad (6)$$

Overall, the proposed self-reflective, difficulty-aware training framework establishes a closed-loop mechanism that dynamically adjusts the training data distribution to the model’s evolving capability. By prioritizing samples based on estimated difficulty and confidence, the framework selects instances that yield effective learning signals under corrupted conditions while filtering low-utility

ones. Although periodic re-evaluation incurs modest computational overhead, this cost is negligible relative to the high per-sample cost of reinforcement learning on videos. In addition, selectively discarding uninformative instances leads to substantial gains in training efficiency, which is discussed in the later section.

3.3 Dual-Branch Alignment Optimization

ROVA trains the model through a dual-branch alignment mechanism that aligns representations from clean and partially perturbed video inputs. The training objective enforces consistency between two branches using the proposed reward modeling combined with GRPO [20]. Here, the clean video branch serves as a fixed anchor with gradients detached, while the perturbed branch is optimized to align its outputs with those of the clean branch. Given a group of G paired samples, the clean branch produces reference outputs $\{o_j\}_{j=1}^G$ and the perturbed branch generates aligned outputs $\{\tilde{o}_j\}_{j=1}^G$. Each pair $(o_j, \tilde{o}_j)$ corresponds to the same video query under clean and perturbed visual conditions. We treat a F_θ as a policy that generates reasoning outputs conditioned on the video inputs:

$$J(\theta) = \mathbb{E}_{(q,V) \sim \mathcal{D}, \{o_j\}_{j=1}^G \sim F_{\theta_{\text{old}}}(O|q,V)} \frac{1}{G} \sum_{j=1}^G [\min(r_j A_j, \text{clip}(r_j, 1 - \epsilon, 1 + \epsilon) A_j) - \beta D_{\text{KL}}(F_\theta \| F_{\text{ref}})], \quad (7)$$

where $r_j = F_\theta(o_j|q)/F_{\text{old}}(o_j|q)$, ϵ and β are hyperparameters, and $D_{\text{KL}}(F_\theta \| F_{\text{ref}})$ denotes the KL-divergence penalty term. The advantage A_j corresponding to output o_j is calculated from the associated reward set $\{r_1, r_2, \dots, r_G\}$:

$$A_j = \frac{r_j - \text{mean}(\{r_1, r_2, \dots, r_G\})}{\text{std}(\{r_1, r_2, \dots, r_G\})}. \quad (8)$$

Format Reward. The model is required to generate an output o_j consisting of an embodied reasoning process p_j followed by a final answer a_j , enclosed within `<think></think>` and `<answer></answer>` tags, respectively. Compliance with this format is verified via a regular expression, producing the format reward r_j^{F} .

$$r_j^{\text{F}} = \begin{cases} 1, & \text{if the format is correct;} \\ 0, & \text{if the format is incorrect.} \end{cases} \quad (9)$$

Accuracy Reward. The accuracy reward r_j^{Acc} evaluates whether the extracted answer o_j is semantically consistent with the ground truth g . Multiple-choice questions typically have a unique and precise answer that can be directly compared once the response follows the required format.

$$r_j^{\text{Acc}} = \begin{cases} 1, & o_j = g; \\ 0, & o_j \neq g. \end{cases} \quad (10)$$

Alignment Reward. For each output pair $(o_j, \tilde{o}_j)$, the alignment reward is decomposed into reasoning and answer components: $r_j^A = r_j^{\text{align}, r} + r_j^{\text{align}, a}$, where $r_j^{\text{align}, r} = \alpha_r \cdot \text{Sim}^r(o_j, \tilde{o}_j)$ and $r_j^{\text{align}, a} = \alpha_a \cdot \text{Sim}^a(o_j, \tilde{o}_j)$. Here, α_r and α_a weight the respective contributions, with Sim^r and Sim^a to measure semantic consistency in the reasoning process and answer segment (see prompt templates in the appendix). The total reward combines alignment with three rewards: $R_j = r_j^F + r_j^{\text{Acc}} + r_j^A$.

With the proposed dual-branch alignment framework, the model is optimized via GRPO using a combined reward signal, including robustness-aware consistency, encouraging stable reasoning and answer predictions across clean and perturbed video inputs, thereby improving robustness and generalization.

4 PVRBench: Evaluating Video Reasoning under Various Realistic Disturbances

Motivation. Existing video reasoning benchmarks, including MVBench [12], Video-MME [8], ALFRED [22], Ego4D [9], and UrbanVideo [40], evaluate models primarily under clean visual conditions (Tab. 1). In contrast, real-world deployment often exposes VLMs to adverse weather, dynamic occlusions, abrupt illumination changes, and camera instability. As shown in Tab. 1, such perturbations can degrade both accuracy and reasoning quality by 12 to 35%. Although ImageNet-C [29] introduced the evaluation of corruption robustness for image classification, no existing benchmark systematically measures how temporally coherent and spatially grounded visual perturbations affect reasoning over videos. This leaves a critical blind spot: we lack the tools to diagnose whether failures under visual corruption arise from perceptual errors, reasoning fragility, or both.

Construction. To close this gap, we introduce *Perturbed Video Reasoning Benchmark (PVRBench)*, designed to evaluate the robustness of video reasoning models under structured, real-world visual variations beyond simple pixel-level corruption. Our focus is on *reasoning reliability*, defined as the ability to maintain coherent and logically consistent inference chains grounded in correct visual observations and valid causal steps despite degraded video input. PVRBench integrates four categories of realistic, video-specific disturbances: lighting (dusk, night, overexposure, shadow), camera motion (translation, zoom, rotation), occlusion (static, dynamic), and weather (fog, rain, snow). Each disturbance is applied with spatial awareness (e.g., depth-conditioned occlusion placement and scene-adapted weather rendering) and temporal coherence across frames. The

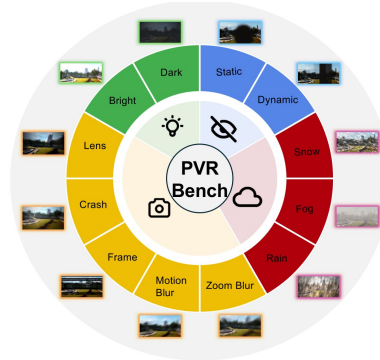


Fig. 3: Overview of the perturbation types in PVRBench.

benchmark comprises over 9K videos and 51K question-answer pairs spanning diverse indoor, outdoor, and embodied scenarios, with 27 task coverage from [30,40], which exercise a broad spectrum of video reasoning capabilities.

Perturbation Injection. At its core, we generate *video-specific masks* (Eq. (2)) that contain semantically coherent perturbations conditioned on each video’s content, including depth layout, object locations, and motion patterns. These perturbations are contextually adapted to scene semantics; for instance, weather appears as windshield rain refraction in driving scenes, while occlusions are placed at plausible foreground locations. For benchmark evaluation, we adopt a static protocol in which masks are pre-generated and fixed per video to ensure reproducible cross-model comparison, while ROVA training (Sec. 3) uses a dynamic protocol that generates perturbations on the fly with stochastically sampled parameters at each iteration to prevent overfitting and promote perturbation invariant representation learning.

Evaluation Metrics. To quantify reasoning reliability, PVRBench introduces five complementary metrics (Fragility, Consistency, Belief, Recovery, and Attention; see Tab. 2) that assess the quality and stability of intermediate reasoning, as well as final-answer accuracy. To assess reasoning process quality, we leverage a powerful vision-language foundational model (e.g., GPT-4o) to score reasoning traces in coherence, perturbation awareness, and evidence grounding via a structured template (see prompt templates in the appendix), following the LLM-as-judge paradigm [10,44].

5 Experiment

5.1 Implementation Details.

Models. We train our model on 4 NVIDIA A100 (80GB) GPUs. For optimization, we set the ordered group size to $G = 8$ and the shuffled group size to $\tilde{G} = G/2$. Details are provided in the appendix.

Datasets. We use both clean and perturbed video data for training and evaluation. For training, we curate an outdoor-scene-relevant subset of Video-R1-260k ($\sim 10\%$ of its video split, filtered by scene category labels) and apply dynamic, randomly sampled perturbation masks to construct corruption-augmented video-query pairs. For evaluation, we assess generalization on the proposed PVRBench, which contains over 51K question answer pairs across more than 9K videos spanning diverse scene categories beyond the training distribution. Static perturbation masks are systematically injected to measure model accuracy, reasoning quality, and robustness under both clean and corrupted conditions. We further evaluate the generalization of VLMs on standard VisBench and UrbanVideo.

5.2 Main Results

ROVA Performance on PVRBench. We extensively evaluate our approach on PVRBench and the clean benchmark (Orig.: UrbanVideo and VSI-Bench)

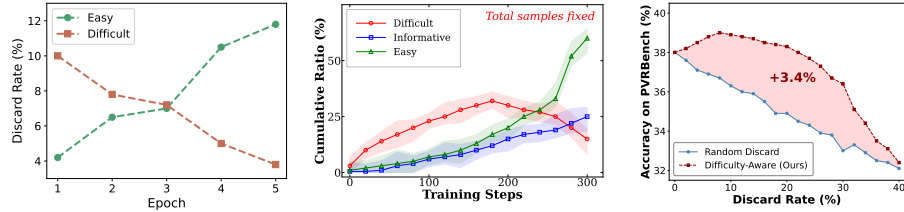
Table 2: Evaluation on PVRBench. We report accuracy under four visual perturbations (**L**ighting, **O**ccclusion, camera **S**hake, **W**eather) on the left, and reasoning quality metrics on the right, including **F**ragility, **C**onsistency, **B**elief, **R**ecovery, and **A**ttention (0 - 5 scale; Higher is better, except for *Fra* ($\downarrow$)). *#Fr*: the number of frames, *Avg.*: the average performance, and *Orig.*: the average performance on clean (unperturbed) data. We exclude *Fra.* when computing *Avg.*[†] and *Orig.*[†].

Model	Size	#Fr	Answer Accuracy							Reasoning Quality							
			Lig.	Occ.	Sha.	Wea.	Avg.	Orig.	Fra.↓	Con.	Bel.	Rec.	Att.	Avg. [†]	Orig. [†]		
Proprietary Models																	
GPT-4o	–	32	.54	.47	.50	.52	.51 _{↓14%}	.59	1.85	3.42	3.55	3.38	3.21	3.39 _{↓11%}	3.82		
Gemini-3-Pro	–	32	.57	.52	.54	.55	.55 _{↓11%}	.62	1.72	3.61	3.48	3.58	3.41	3.52 _{↓10%}	3.91		
Claude-3.5-Son.	–	32	.45	.41	.44	.45	.44 _{↓17%}	.53	2.08	3.18	3.22	2.95	3.15	3.13 _{↓14%}	3.65		
Video Reasoning Models																	
Video-R1	7B	32	.43	.37	.42	.41	.41 _{↓20%}	.51	2.48	2.75	2.85	2.68	2.65	2.73 _{↓20%}	3.42		
Video-R1	72B	32	.51	.45	.49	.49	.49 _{↓16%}	.58	2.11	3.25	3.18	3.21	2.98	3.16 _{↓14%}	3.68		
VideoChat-R	7B	16	.36	.31	.36	.35	.35 _{↓22%}	.45	2.65	2.62	2.55	2.71	2.28	2.54 _{↓22%}	3.25		
LLaVA-Video-R	7B	32	.40	.34	.38	.38	.38 _{↓21%}	.48	2.58	2.68	2.61	2.78	2.42	2.62 _{↓21%}	3.32		
Embodied-R	7B	32	.45	.38	.42	.43	.42 _{↓22%}	.54	2.45	2.82	2.91	2.72	2.68	2.78 _{↓19%}	3.45		
† + ROVA (Ours)	7B	32	.52	.46	.49	.51	.50 _{↑9%}	.55	2.25	3.15	3.18	3.22	2.91	3.12 _{↑13%}	3.58		
Open-Source Video LLMs																	
LLaVA-Video	7B	32	.32	.29	.30	.32	.31 _{↓30%}	.44	2.78	2.45	2.35	2.52	2.25	2.39 _{↓23%}	3.12		
VideoLLaMA2	7B	16	.28	.25	.27	.29	.27 _{↓25%}	.36	2.92	2.18	2.25	2.12	2.15	2.18 _{↓28%}	3.01		
VideoChat2	7B	16	.26	.23	.25	.27	.25 _{↓26%}	.34	3.01	2.08	2.15	2.05	2.02	2.08 _{↓28%}	2.88		
MiniCPM-V 2.6	8B	64	.34	.28	.31	.32	.31 _{↓28%}	.43	2.75	2.48	2.42	2.55	2.21	2.42 _{↓24%}	3.18		
InternVL2.5	8B	32	.31	.26	.32	.33	.31 _{↓33%}	.46	2.85	2.38	2.28	2.42	2.18	2.32 _{↓26%}	3.15		
† + ROVA (Ours)	8B	32	.43	.36	.41	.40	.40 _{↑15%}	.47	2.45	2.82	2.75	2.78	2.58	2.73 _{↑17%}	3.28		
Qwen2.5-VL	7B	32	.35	.28	.34	.34	.33 _{↓35%}	.51	2.71	2.58	2.62	2.68	2.31	2.55 _{↓25%}	3.41		
† + ROVA (Ours)	7B	32	.48	.43	.47	.49	.47 _{↑11%}	.53	2.31	3.05	3.08	2.98	2.85	2.99 _{↑15%}	3.52		
Qwen2.5-VL	72B	32	.48	.41	.44	.47	.45 _{↓21%}	.57	2.18	3.15	3.08	2.92	3.12	3.07 _{↓16%}	3.64		
† + ROVA (Ours)	72B	32	.57	.53	.56	.56	.56 _{↑5%}	.59	1.95	3.45	3.35	3.42	3.18	3.35 _{↑10%}	3.72		
Qwen3-VL	13B	32	.43	.35	.39	.42	.40 _{↓25%}	.53	2.41	2.85	2.92	2.78	2.72	2.82 _{↓10%}	3.48		
† + ROVA (Ours)	13B	32	.53	.49	.52	.54	.52 _{↑7%}	.56	2.12	3.28	3.32	3.18	3.05	3.21 _{↑11%}	3.62		

across diverse backbones, including video reasoning models and open-source video LLMs ranging from 7B to 72B. As shown in Tab. 2, among dedicated video reasoning models, ROVA consistently outperforms prior methods. In the 7B setting, it improves the best-performing model, Embodied-R, from 0.42 to 0.50 average accuracy under perturbations (more than 17% relative gain), and even matches or surpasses the much larger Video-R1 72B. Importantly, it also achieves consistent improvements in reasoning quality, indicating stable and reliable reasoning under visual corruption. Most open-source video LLMs suffer substantial degradation under perturbations, with 21–35% drops in accuracy and 16–28% declines in reasoning quality relative to clean inputs. Notably, ROVA not only withstands the proposed perturbations but also enhances the model’s generalization performance, observing consistent gains on PVRBench and across unseen benchmarks (VisBench and UrbanVideo) in both answer accuracy and reasoning quality under clean and perturbed videos. These findings suggest that ROVA is able to learn perturbation-robust representations with strong transferability, enabling improved robustness and semantic understanding beyond

Table 3: Training efficiency comparison (Qwen2.5-VL-7B, Orig. Acc. = 0.43; GPU-h = #GPUs × wall-clock hours). SRE = Self-Reflective Evaluation, DRE = Difficulty Re-Evaluation, ME = Memory Eviction. Robust. = dual-branch alignment with structured corruption (Secs. 3.1 and 3.3). Curric. = SRE + DRE + ME (Sec. 3.2).

	Training Data			Architecture			Config.	Performance	
	SFT	RL	Total	Branch	Robust.	Curric.	GPUs	GPU-h	Avg. Acc.
Std. GRPO	—	—	—	Single	✗	✗	4×A100	71.6	.45
Naïve Dual	—	—	—	Dual	✓	✗	4×A100	142.8	.48
Video-R1	165K	260K	425K	Single	✗	✗	8×A100	339.2	.49
ROVA	6.5K	26K	32.5K	Dual	✓	✓	4×A100	134.4	.53



(a) Sample discard rate evolution during self-reflective curriculum training. (b) Evolution of estimated easy, informative, and difficult sample proportions over training. (c) Difficulty-aware confidence-threshold discard vs. random across retention.

Fig. 4: Analysis of Self-Reflective Evaluation and Difficulty-Aware Training for ROVA during the first Epoch of Qwen-VL-2.5-7B Training.

the training distribution without domain-specific fine-tuning, while maintaining superior performance on clean data. Beyond the accuracy and reasoning quality improvements, Tab. 3 shows that ROVA is highly resource-efficient. Although the dual-branch design doubles the forward pass, the proposed curriculum (SRE + DRE + ME) more than offsets this overhead, reducing GPU-hours by 5.9% compared to naive Dual-Branch (134.4 vs. 142.8) while improving accuracy from 0.37 to 0.47. Moreover, ROVA surpasses Video-R1 by 23.7% (0.47 vs. 0.38) while using 60.4% fewer GPU-hours (134.4 vs. 339.2), half the GPUs, and less than 8% of the training data (32.5K vs. 425K). These results suggest that the dual-branch alignment objective learns transferable, perturbation-robust representations that generalize beyond the training distribution without domain-specific fine-tuning, while maintaining strong performance on clean data.

Analysis of self-reflective evaluation and sample-selective training. We also analyze the behavior of our self-reflection evaluation mechanism during training. As shown in Fig. 4a, the discard rate for easy samples increases steadily over epochs while that for difficult samples declines, indicating that the model keeps evolving and smarter and prefers to decline more samples as they are already good at those, Fig. 4a, shows a moderate fraction of samples is discarded overall, and the model selectively filters low-utility or overly noisy instances rather than aggressively pruning data. Fig. 4b further illustrates the evolution of the estimated sample difficulty in training steps. While the total number of discarded samples is fixed, the composition gradually shifts toward easy samples, reflecting the improving competence of the model: samples initially deemed difficult are

Table 4: We compare with five robustness baselines on PVRBench, where the base model (QwenVL-2.5-7B) uses 16 frames.

Method	Venue	Accuracy	Reasoning	GPU-hours
QwenVL-2.5-7B	–	0.33	2.55	–
MVTamperBench [1]	ACL 2025	0.38	2.68	86.4
Dbl. Visual Def. [44]	WACV 2025	0.40	2.74	108.2
PanoDPO [12]	ACL 2025	0.43	2.88	146.0
Ro-Bench [32]	ICASSP 2026	0.42	2.83	94.6
ROVA (Ours)	-	0.47	2.99	134.4

Table 5: Generalization to unseen perturbations on MSRVT-P and YouCook2-P (“P” stands for Perturbations).

Benchmark	Average Accuracy		Average Reasoning	
	Base	ROVA (Ours)	Base	ROVA (Ours)
MSRVT-P	0.38	0.46 (+21%)	2.41	2.88 (+19%)
YouCook2-P	0.35	0.43 (+23%)	2.32	2.79 (+20%)

increasingly reclassified as easy as training progresses. This dynamic redistribution suggests that the self-reflective evaluator captures meaningful learning signals and adapts the curriculum in a data-driven manner. Fig. 4c demonstrates the effectiveness of difficulty-aware data selection for training. Compared to random discarding, our strategy consistently achieves higher accuracy across discard rates, with an improvement of up to 3.4% on PVRBench. This indicates that selective removal of samples based on estimated difficulty preserves informative training signals while avoiding detrimental noise.

Analysis of Performance and Robustness to Perturbations. To thoroughly evaluate ROVA, we compare it with five representative robustness baselines on PVRBench (Tab. 4). Using the same base model (QwenVL-2.5-7B), ROVA achieves the best overall performance, with an accuracy of 0.47 and a reasoning score of 2.99, substantially outperforming MVTamperBench (0.38/2.68) and PanoDPO (0.43/2.88). Notably, ROVA yields consistent gains even with lower training cost (134.4 vs. 146.0 GPU-hours for PanoDPO), showing a favorable efficiency–effectiveness trade-off. For cross-dataset generalisation on MSRVT-P and YouCook2-P with unseen perturbations (Tab. 5), the base model drops significantly, while ROVA demonstrates strong resilience: compared to the base model, it improves average accuracy by 21% and 23%, and average reasoning by 19% and 20%, respectively. These consistent gains, on both in-distribution and out-of-distribution perturbations, indicate that ROVA does not overfit to a specific training distribution, but learns more discriminative and robust multimodal representations.

5.3 Ablation Study and Analysis

Ablation of Core Components. We ablate each component of ROVA to assess its contribution (Fig. 5a). The reasoning reward yields the largest gain, followed

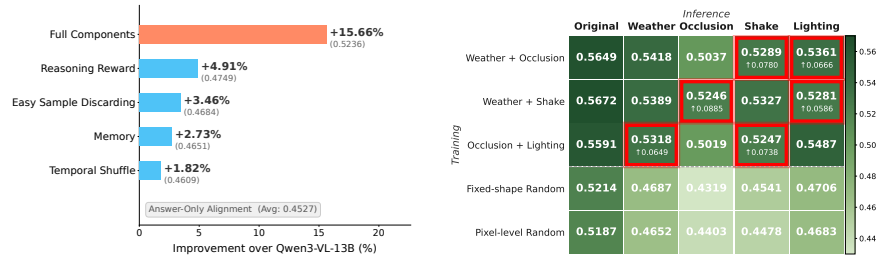


Fig. 5: Ablation studies of ROVA. (a) Impact of individual components on answer accuracy. (b) Comparison of corruption mask strategies across perturbation types. Experiments are conducted using the Qwen3-VL-13B model trained for 3 epochs.

by easy sample discarding, underscoring the central role of semantic reasoning and targeted curation. The memory module and temporal shuffle provide smaller but consistent gains, serving as complementary regularizers that stabilize training and enhance robustness.

Ablation of Mask Styles. We explore the generalizability of the proposed structured masking strategy compared to random masking baselines. As shown in Fig. 5b, models trained on only two corruption mask styles achieve strong in-domain performance on the perturbation types seen during training, and more importantly, transfer effectively to held-out perturbation types (highlighted in red): out-of-domain performance remains close to in-domain results while both consistently surpass fixed-shape and pixel-level random masking by a significant margin (6 - 9% absolute). This indicates that structured, perturbation-aware masks capture transferable corruption patterns rather than overfitting to specific disturbance types, confirming that a small subset of mask styles suffices to achieve broad robustness under diverse real-world disturbances.

Ablation of reward models. Notably, our LLM judge (GPT-4o by default) outperforms rule- or embedding-based matching in evaluating semantic consistency across reasoning traces and final answers. Replacing it with open-source models (e.g., Qwen3-13B) yields comparable results, suggesting that the approach generalizes beyond proprietary APIs (Tab. 6). In contrast, more granular reward designs, such as conditional alignment or step-level consistency, introduce additional variance that destabilizes GRPO and degrades performance, further supporting LLM-based evaluation as the most effective approach.

Table 6: Ablation study of the reward model on PVRBench using commercial and open source VLMs.

Reward Judge	Acc.	Avg. [†]	Free
GPT-4o	.470	2.99	✗
Qwen3-13B	.467	2.97	✓
Qwen2.5-7B	.463	2.95	✓



Fig. 6: Qualitative examples of ROVA-trained Qwen2.5-VL-7B performing obstacle avoidance and target identification under night-time low-light conditions.

5.4 Qualitative Analysis

We further validate the robustness of ROVA through qualitative examples on representative tasks in Fig. 6. Even in challenging scenarios where adverse weather or visual disturbances significantly degrade visibility, ROVA remains effective, correctly reasoning about the scene and task requirements. For instance, when heavy rain and glare obscure key visual cues, ROVA can still infer spatial relationships and scene structure, and when large objects block the field of view, it correctly reasons about the underlying layout rather than relying on partial appearances. This shows that ROVA reliably interprets and reasons in visually impaired conditions, demonstrating robustness beyond controlled settings and confirming its effectiveness in difficult, realistic environments.

6 Conclusion

In this work, we present ROVA, a robust training framework for embodied video reasoning that leverages structured spatio-temporal corruptions, dual-branch alignment, and self-reflective data curation to learn perturbation-robust representations. To evaluate robustness under realistic disturbances, we introduce PVRBench. We show that ROVA consistently improves robustness under diverse real-world perturbations in video inputs while also improving performance on clean video-question pairs. These contributions provide both a principled benchmark and a practical training recipe, enabling future studies on broader perturbation families and more complex long-horizon embodied tasks.

References

1. Agarwal, A., Panda, S., Charles, A., Patel, H.L., Kumar, B., Pattnayak, P., Rafi, T.H., Kumar, T., Meghwani, H., Gupta, K., Chae, D.K.: MVTamperBench: Evaluating robustness of vision-language models. In: Findings of the Association for Computational Linguistics: ACL 2025. pp. 18804–18828 (2025)
2. Azzolini, A., Bai, J., Brandon, H., Cao, J., Chattopadhyay, P., Chen, H., Chu, J., Cui, Y., Diamond, J., Ding, Y., et al.: Cosmos-reason1: From physical common sense to embodied reasoning. arXiv preprint arXiv:2503.15558 (2025)
3. Cai, Y., Liu, Y., Zhang, Z., Shi, J.Q.: Clap: Isolating content from style through contrastive learning with augmented prompts. In: ECCV2024 (2024)
4. Chen, R., Fan, Z., Luo, T., Zou, H., Feng, Z., Xie, G., Zhang, H., Wang, Z., Liu, Z., Zhang, H.: Datasets and recipes for video temporal grounding via reinforcement learning. In: Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track. pp. 983–992 (2025)
5. Chen, S., Wu, Z., Zhang, K., Li, C., Zhang, B., Ma, F., Yu, F.R., Li, Q.: Exploring embodied multimodal large models: Development, datasets, and future directions. arXiv preprint arXiv:2502.15336 (2025)
6. Clark, C., Zhang, J., Ma, Z., Park, J.S., Salehi, M., Tripathi, R., Lee, S., Ren, Z., Kim, C.D., Yang, Y., et al.: Molmo2: Open weights and data for vision-language models with video understanding and grounding. arXiv preprint arXiv:2601.10611 (2026)
7. Duan, J., Fan, Q., Cheng, H., Shi, X., Xu, K.: Improve video representation with temporal adversarial augmentation. In: Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence. pp. 708–716 (2023)
8. 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 IEEE/CVF conference on computer vision and pattern recognition. pp. 24108–24118 (2025)
9. Grauman, K., Westbury, A., Byrne, E., Chavis, Z., Furnari, A., Girdhar, R., Hamburger, J., Jiang, H., Liu, M., Liu, X., et al.: Ego4d: Around the world in 3,000 hours of egocentric video. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 18995–19012 (2022)
10. He, X., Jiang, D., Zhang, G., Ku, M., Soni, A., Siu, S., Chen, H., Chandra, A., Jiang, Z., Arulraj, A., et al.: Videoscore: Building automatic metrics to simulate fine-grained human feedback for video generation. In: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. pp. 2105–2123 (2024)
11. 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)
12. Li, K., et al.: Mvbench: A comprehensive multi-modal video understanding benchmark. In: CVPR (2024)
13. 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)
14. Maaz, M., Rasheed, H., Khan, S., Khan, F.: Video-chatgpt: Towards detailed video understanding via large vision and language models. In: Proceedings of the Association for Computational Linguistics (ACL). pp. 12585–12602 (2024)

15. Mao, C., Geng, S., Yang, J., Wang, X., Vondrick, C.: Understanding zero-shot adversarial robustness for large-scale models. *arXiv preprint arXiv:2212.07016* (2022)
16. Mao, C., Geng, S., Yang, J., Wang, X., Vondrick, C.: Understanding zero-shot adversarial robustness for large-scale models. In: *ICLR2023* (2023)
17. Nguyen, T., Bin, Y., Xiao, J., Qu, L., Li, Y., Wu, J.Z., Nguyen, C.D., Ng, S.K., Luu, A.T.: Video-language understanding: A survey from model architecture, model training, and data perspectives. In: *Findings of the Association for Computational Linguistics: ACL 2024*. pp. 3636–3657 (Aug 2024)
18. Oh, C., Fang, Z., Im, S., Du, X., Li, Y.: Understanding multimodal LLMs under distribution shifts: An information-theoretic approach. In: *Forty-second International Conference on Machine Learning* (2025)
19. Schiappa, M.C., Vyas, S., Palangi, H., Rawat, Y.S., Vineet, V.: Robustness analysis of video-language models against visual and language perturbations. In: *36th Conference on Neural Information Processing Systems (NeurIPS 2022) Track on Datasets and Benchmarks* (2022)
20. Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Bi, X., Zhang, H., Zhang, M., Li, Y., Wu, Y., et al.: Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300* (2024)
21. Sheng, L., Liang, J., Wang, Z., He, R.: R-tpt: Improving adversarial robustness of vision-language models through test-time prompt tuning. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 29958–29967 (2025)
22. Shridhar, M., Thomason, J., Gordon, D., Bisk, Y., Han, W., Mottaghi, R., Zettlemoyer, L., Fox, D.: Alfred: A benchmark for interpreting grounded instructions for everyday tasks. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 10740–10749 (2020)
23. Shu, Y., Liu, Z., Zhang, P., Qin, M., Zhou, J., Liang, Z., Huang, T., Zhao, B.: Video-xl: Extra-long vision language model for hour-scale video understanding. In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 26160–26169 (2025)
24. Team, G., Anil, R., Borgeaud, S., Alayrac, J.B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A.M., Hauth, A., Millican, K., et al.: Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805* (2023)
25. Tong, B., Lai, H., Pan, Y., Yin, J.: On the zero-shot adversarial robustness of vision-language models: A truly zero-shot and training-free approach. In: *Proceedings of the Computer Vision and Pattern Recognition Conference*. pp. 19921–19930 (2025)
26. Wang, Y., Wang, Z., Xu, B., Du, Y., Lin, K., Xiao, Z., Yue, Z., Ju, J., Zhang, L., Yang, D., Fang, X., He, Z., Luo, Z., Wang, W., Lin, J., Luan, J., Jin, Q.: Time-r1: Post-training large vision language model for temporal video grounding. In: *39th Conference on Neural Information Processing Systems (NeurIPS 2025)*. (2025)
27. 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)
28. Wu, Y., Schlegel, V., Batista-Navarro, R.: Pay attention to real world perturbations! natural robustness evaluation in machine reading comprehension. *arXiv preprint arXiv:2502.16523* (2025)
29. Xie, Q., Luong, M.T., Hovy, E., Le, Q.V.: Self-training with noisy student improves imagenet classification. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 10687–10698 (2020)

30. Yang, J., Yang, S., Gupta, A.W., Han, R., Fei-Fei, L., Xie, S.: Thinking in space: How multimodal large language models see, remember, and recall spaces. In: *Proceedings of the Computer Vision and Pattern Recognition Conference*. pp. 10632–10643 (2025)
31. Yang, Z., Li, J., Diao, M., Jing, Y., Liang, K.: Ro-bench: Large-scale robustness evaluation of mllms with text-driven counterfactual videos. *arXiv:2510.08936* (2025)
32. Yeo, W., Kim, K., Yoon, J., Hwang, S.J.: Worldmm: Dynamic multimodal memory agent for long video reasoning. In: *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)* (2026)
33. Yu, S., Yoon, J., Bansal, M.: Crema: Generalizable and efficient video-language reasoning via multimodal modular fusion. In: *Proceedings of the International Conference on Learning Representations (ICLR)* (2025)
34. 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)
35. Yuan, L., Wang, J., Sun, H., Zhang, Y., Lin, Y.: Tarsier2: Advancing large vision-language models from detailed video description to comprehensive video understanding. *arXiv preprint arXiv:2501.07888* (2025)
36. Zeng, R., Chen, X., Liang, J., Wu, H., Cao, G., Guo, Y.: Benchmarking the robustness of temporal action detection models against temporal corruptions. In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 18263–18274 (2024)
37. Zhang, H., Li, X., Bing, L.: Video-llama: An instruction-tuned audio-visual language model for video understanding. In: *Proceedings of the 2023 conference on empirical methods in natural language processing: system demonstrations*. pp. 543–553 (2023)
38. Zhang, J., Pang, T., Du, C., Ren, Y., Li, B., Lin, M.: Benchmarking large multimodal models against common corruptions. *arXiv preprint arXiv:2401.11943* (2024)
39. Zhang, W., Wang, M., Liu, G., Huixin, X., Jiang, Y., Shen, Y., Hou, G., Zheng, Z., Zhang, H., Li, X., et al.: Embodied-reasoner: Synergizing visual search, reasoning, and action for embodied interactive tasks. *arXiv preprint arXiv:2503.21696* (2025)
40. Zhao, B., Fang, J., Dai, Z., Wang, Z., Zha, J., Zhang, W., Gao, C., Wang, Y., Cui, J., Chen, X., et al.: Urbanvideo-bench: Benchmarking vision-language models on embodied intelligence with video data in urban spaces. In: *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. pp. 32400–32423 (2025)
41. Zhao, B., Wang, Z., Fang, J., Gao, C., Man, F., Cui, J., Wang, X., Chen, X., Li, Y., Zhu, W.: Embodied-r: Collaborative framework for activating embodied spatial reasoning in foundation models via reinforcement learning. In: *Proceedings of the 33rd ACM International Conference on Multimedia* (2025)
42. Zhao, S., Wang, X., Zhu, L., Yang, Y.: Test-time adaptation with clip reward for zero-shot generalization in vision-language models. In: *ICLR2024* (2024)
43. Zhao, Y., Pang, T., Du, C., Yang, X., Li, C., Cheung, N.M., Lin, M.: On evaluating adversarial robustness of large vision-language models. In: *37th Conference on Neural Information Processing Systems (NeurIPS 2023)*. (2023)
44. Zheng, L., Chiang, W.L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E., et al.: Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems* **36**, 46595–46623 (2023)
45. Zhou, W., Bai, S., Mandic, D.P., Zhao, Q., Chen, B.: Revisiting the adversarial robustness of vision language models: a multimodal perspective. *arXiv preprint arXiv:2404.19287* (2024)

46. Zhou, Y., He, Y., Su, Y., Han, S., Jang, J., Bertasius, G., Bansal, M., Yao, H.: Reagent-v: A reward-driven multi-agent framework for video understanding. In: 39th Conference on Neural Information Processing Systems (NeurIPS 2025). (2025)