RSVideo: Are Your Vision-Language Models Ready for Remote Sensing Videos?

Hongjie Zhou[1,2,*] Shiqin Wang[1,*] Haoyang Chen[1,2,*] Haonan Guo[1,2]
Di Wang[1,2,†] Juhua Liu[1] Fu Lin[1] Yong Luo[1,†]
[1]Wuhan University [2]Zhongguancun Academy
[*]Equal contribution. [†]Corresponding authors.
d_wang@whu.edu.cn, luoyong@whu.edu.cn

Abstract

*Remote-sensing videos enable real-time observation of changes in target attributes, short-term activities, and scene evolution. They record motion, actions, interactions, and scene changes that cannot be captured by isolated images. Existing models primarily target single images or discrete temporal observations spanning a long time range. However, a unified evaluation setting for assessing vision-language models on continuous remote-sensing video understanding remains lacking. We introduce **RSVideo-10K**, a remote-sensing video dataset comprising 10,773 instances, 1.47 million frames, and 17.02 hours of footage, containing both unmanned aerial vehicles and satellite platforms. Its fixed evaluation benchmark, **RSVideo-Bench**, contains 2,731 test instances and evaluates two complementary aspects of remote-sensing video understanding: L1 Perception and L2 Reasoning, spanning seven capability groups and 17 tasks. Evaluations show that current vision-language models still struggle to recover small local evidence, track short-lived states, and use scene-constrained spatial relations. Based on this analysis, we further propose **RSVideo**, a reinforcement learning framework for small-target spatiotemporal focusing that selects question-relevant regions across frames and suppresses redundant background tokens. RSVideo achieves a maximum absolute improvement of 9.01% with InternVL3.5-14B and attains the highest accuracy of 40.63% with Qwen3.6-27B across 26 open-source vision-language backbones.*

1. Introduction

Remote-sensing videos are becoming an important source for traffic monitoring, public-safety surveillance, moving-target analysis, disaster response, and environmental observation [4, 6, 23, 26, 49]. Unlike a single remote-sensing image, a video continuously observes the scene and captures transient changes. This enables real-time observation of target attribute changes, short-term activities, and scene evolution. Recent vision-language models improve the automation of video analysis and achieved strong performance on natural-video tasks [3, 13, 47].

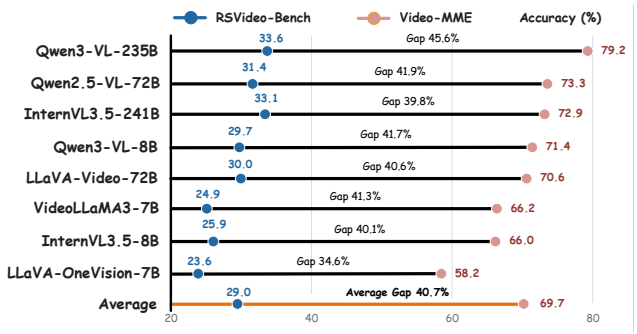


Figure 1. Accuracy of representative vision-language models on RSVideo-Bench (ours) and Video-MME [9]. Existing multimodal large language models achieve an average accuracy of 69.7% on natural videos, but only 29.0% on remote-sensing videos, resulting in an average performance gap of 40.7%.

In remote-sensing videos, targets may occupy only a few pixels, while objects with similar appearances are difficult to distinguish. Meanwhile, critical actions and state changes can be subtle and short-lived, and complex backgrounds introduce substantial redundancy. These challenges lead to a fundamental question: *How well do current vision-language models understand remote-sensing videos?*

Answering this question using existing remote-sensing benchmarks is difficult. Static remote-sensing vision-language datasets support scene understanding, visual question answering, captioning and visual grounding [12, 16, 19, 24, 34, 36]. Their basic unit is a single image or region. They cannot capture continuous motion or short-lived states. Multi-temporal remote-sensing datasets [17, 30, 38, 44] support long-term change analysis. However, their observations are usually separated by intervals of days, months, or years. This distinguishes them from videos,

which capture movements and state changes in seconds. Existing UAV datasets mainly target low-altitude perception, grounding, or task-specific reasoning [4–6, 23, 28, 29, 45, 46]. They are insufficient for evaluating continuous remote-sensing video understanding across UAV and satellite platforms, where target scales and visual evidence evolve over time.

To support both controlled evaluation and model development, we introduce **RSVideo-10K**. To our knowledge, this is the first large-scale remote-sensing video dataset designed for video question answering at the 10K-instance scale. It comprises **RSVideo-Instruct**, which provides 8,042 training and validation instances, and **RSVideo-Bench**, a fixed test set containing 2,731 instances. RSVideo-10K contains continuous videos from eight public sources spanning UAV and satellite observations, with a total duration of approximately 17.02 hours and 1,473,150 decoded frames. All instances are organized under a unified taxonomy covering two capability dimensions and 17 tasks.

We use RSVideo-Bench to evaluate representative open-weight and proprietary vision-language models. It contains 653 L1 Perception instances and 2,078 L2 Reasoning instances. For comparison, we perform evaluations on the natural video dataset Video-MME [9]. All models are evaluated using the same video input and five-choice protocol. As shown in Figure 1, many vision-language models achieve strong performance on Video-MME. However, this performance gain does not directly transfer to remote-sensing videos. Furthermore, we found that their errors concentrate on small targets, visually similar objects, weak or short-lived state changes, and spatial relations defined by roads, buildings, region boundaries, or nearby objects. This pattern points to a specific research question: *How can vision-language models recover question-relevant spatiotemporal evidence and target states from continuous remote-sensing videos?*

RSVideo-Instruct makes these challenges trainable by providing answer supervision together with temporal and spatial evidence annotations, enabling evidence-aware model initialization. Based on this supervision, we propose **RSVideo**, a two-stage training framework for spatiotemporal evidence focusing. The first stage performs evidence-aware initialization, teaching the model to associate predictions with relevant spatiotemporal evidence. The second stage further optimizes this evidence focusing capability through reinforcement learning, enabling adaptive spatiotemporal evidence selection and redundant context compression under a fixed visual token budget. To guide these behaviors, we design evidence-aware rewards that jointly consider answer correctness, spatiotemporal evidence alignment, and background compression. RSVideo achieves consistent performance improvements across multiple open-weight vision-language backbones on RSVideo-Bench. Extensive experiments verify that RSVideo improves remote-sensing video understanding performance across 26 vision-language backbones ranging from 1B to 241B parameters.

The contributions of this work are summarized as follows:

- We introduce **RSVideo-10K**, the first unified dataset for perception and reasoning on continuous UAV and satellite videos. It contains 10,773 instances, including 8,042 training and validation instances from **RSVideo-Instruct** and 2,731 test instances in **RSVideo-Bench**. All instances follow a unified taxonomy spanning two capability dimensions, seven capability groups, and 17 tasks, with expert annotation and review to ensure quality.
- We propose **RSVideo**, a two-stage training framework for spatiotemporal evidence focusing. It leverages temporal and spatial evidence annotations in RSVideo-Instruct for evidence-aware initialization and further optimizes evidence selection through reinforcement learning. Under a fixed visual token budget, RSVideo adaptively preserves informative evidence tokens and compresses redundant background context.
- Extensive experiments demonstrate that existing vision-language models still struggle with evidence-grounded remote-sensing video understanding, particularly for small targets, short-lived states, and scene-constrained spatial relations. RSVideo consistently improves performance across 26 evaluated vision-language backbones, achieving a maximum absolute gain of 9.01% with InternVL3.5-14B.

2. Related Work

2.1. Remote-Sensing Vision-Language Data

Remote-sensing vision-language datasets support question answering, captioning, and visual grounding over satellite and aerial images [12, 16, 19, 34, 36]. XLRS-Bench and UHR-Micro study multi-scale understanding and local evidence extraction from ultra-high-resolution imagery [24, 33]. These datasets provide broad spatial coverage but operate mainly on static images. Multi-temporal datasets instead model changes between observations acquired at different times. DynamicEarthNet and FUSU support land-cover and semantic change analysis, while DynamicVL introduces language-based reasoning over long-term urban changes [30, 38, 44]. However, these observations are typically separated by long intervals and cannot capture continuous motion and short-lived state changes. Recent remote-sensing foundation models have explored multi-task pretraining [31], while geospatial MLLMs have begun to study emergent reasoning without predefined chain-of-thought supervision [32]. In contrast, RSVideo-10K focuses on quality-controlled continuous videos and explicitly eval-

Dataset	Video Scale			Task Taxonomy			Evidence		Splits			Video Coverage	
	Queries	Clips	Hours	Perception	Reasoning	Tasks	Time	Position	Train	Val.	Test	UAV	Satellite
Static Remote-Sensing Vision-Language Datasets													
RSVQA [19]	0	0	0	✓	✗	5	✗	✗	✓	✓	✓	✗	✗
EarthVQA [34]	0	0	0	✓	✓	6	✗	✗	✓	✓	✓	✗	✗
VRSBench [16]	0	0	0	✓	✓	3	✗	✓	✓	✗	✓	✗	✗
XLRS-Bench [33]	0	0	0	✓	✓	16	✗	✓	✗	✗	✓	✗	✗
Multi-Temporal Remote-Sensing Vision-Language Datasets													
TEOChat / TEOChatlas [11]	0	0	0	✓	✓	0	✓	✗	✓	✗	✗	✗	✗
DynamicVL / DVL-Suite [38]	0	0	0	✓	✓	6	✓	✓	✓	✗	✓	✗	✗
VLRS-Bench [20]	0	0	0	✓	✓	14	✓	✗	✗	✗	✓	✗	✗
General Vision-Language Datasets													
VSI-Bench [39]	5,130	288	7.73	✓	✓	8	✗	✓	✗	✗	✓	✗	✗
TempCompass [18]	7,540	410	1.30	✓	✗	4	✗	✗	✗	✗	✓	✗	✗
MVBench [14]	4,000	3,641	13.22	✓	✓	20	✓	✗	✗	✗	✓	✗	✗
Video-MME [9]	2,700	900	255.32	✓	✓	12	✗	✗	✗	✗	✓	✗	✗
UAV-Centric Multimodal Datasets													
UAVBench / UAVIT-1M [46]	0	0	0	✓	✓	10/11	✗	✓	✓	✗	✓	✗	✗
UAVReason [28]	0	0	0	✓	✓	22	✓	✓	✓	✗	✓	✗	✗
UrbanVideo-Bench [48]	5,355	1,393	48.71	✓	✓	16	✗	✗	✓	✗	✓	✓	✗
SIS-Bench [52]	4,856	1,646	14.06	✓	✓	13	✗	✗	✗	✗	✓	✓	✗
RSVideo-10K (ours)	10,773	4,629	17.02	✓	✓	17	✓	✓	✓	✓	✓	✓	✓

Table 1. Comparison with representative vision-language datasets. Queries and Clips denote the number of continuous-video QA instances and unique video inputs, respectively. Hours indicate the total duration of deduplicated video segments used for question construction; each segment is counted once regardless of the number of associated queries. Task Taxonomy denotes explicit capability dimensions and task numbers. Evidence indicates whether temporal or spatial evidence annotations are provided. Splits follow the corresponding papers or official releases. Video Coverage indicates the source platforms of the videos. "0" in Video Scale denotes image-only datasets.

uates fine-grained perception and cross-frame reasoning.

2.2. General Video-Language Understanding

Vision-language datasets have been developed for action recognition, event understanding, temporal reasoning, and long-form video comprehension [9, 14, 22]. Recent methods improve video reasoning through hierarchical memory, adaptive frame selection, active perception, and evidence retrieval [1, 10, 21, 37, 40, 42, 51]. These studies establish strong foundations for video-level reasoning, but are mainly developed for natural videos with relatively salient targets and temporal changes. In contrast, remote-sensing videos contain small targets, large repetitive backgrounds, subtle state changes, and spatial relations constrained by scene structures such as roads, buildings, and region boundaries. These characteristics require models to recover weak and sparse spatiotemporal evidence before performing reliable video reasoning.

2.3. Aerial and Remote-Sensing Video Data

Existing aerial video datasets have primarily focused on task-specific problems, such as object detection, tracking, action recognition, and event analysis [7, 41, 49]. Recent efforts extend aerial video understanding toward vision-language tasks. For example, UAVBench and UAVIT-1M explore UAV vision-language understanding at different granularities, while UAVReason investigates geometric and topological reasoning in simulated nadir-view scenarios [28, 46]. Other benchmarks, such as UrbanVideo-Bench and SIS-Bench, evaluate broader UAV video understanding beyond perception tasks [48, 52]. However, existing aerial video resources remain fragmented across platforms and tasks, lacking comprehensive evaluation of continuous remote-sensing video understanding. In contrast, RSVideo-10K provides a unified framework that integrates continuous UAV and satellite videos with fine-grained perception and reasoning tasks under a shared taxonomy.

3. Remote Sensing Video Understanding Dataset RSVideo-10K

RSVideo-10K contains 10,773 QA instances derived from 4,629 manually audited video clips, covering 17.02 hours of continuous video and 1.47M decoded frames. Each instance contains a video, a question, and five answer choices, with all questions formulated as single-choice questions. Notably, 4,638 instances of RSVideo-10K were additionally annotated temporal and spatial evidence, indicating where and when the question-relevant target or event is located. RSVideo-10K consists of 8,042 **RSVideo-Instruct** instances and 2,731 **RSVideo-Bench** instances under a shared taxonomy and annotation protocol. RSVideo-Instruct provides training and validation data for model development. RSVideo-Bench is a locked test set for evaluating remote-sensing video perception and reasoning. Table 1 compares RSVideo-10K with representative vision-language and remote-sensing datasets.

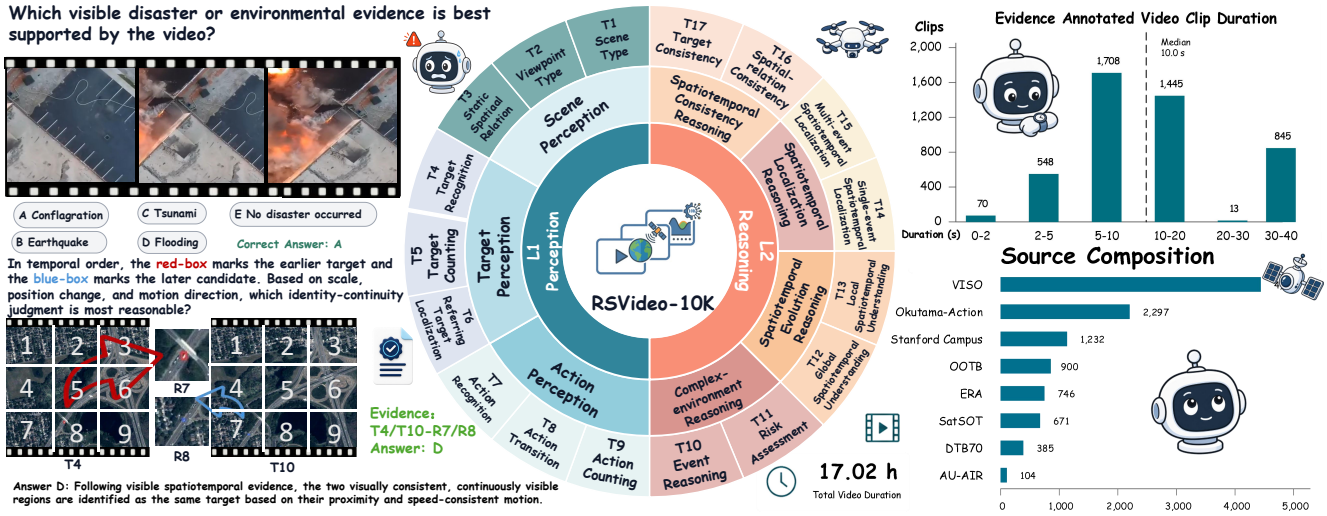


Figure 2. **Overview of RSVideo-10K**. The dataset contains 10,773 five-choice QA instances grounded in 4,629 audited evidence clips from eight public video sources. It covers 17.02 hours of continuous remote-sensing video with 1,473,150 decoded frames. RSVideo-10K comprises RSVideo-Instruct for training and validation and RSVideo-Bench for fixed evaluation. The figure summarizes the source composition, clip duration of evidence annotated videos, and the L1/L2 capability distribution.

3.1. Task Taxonomy

We organize RSVideo-10K with a four-level taxonomy: two capability dimensions, seven capability groups, 17 tasks, and 26 fine-grained leaves. The leaves specify distinct evidence patterns within each task. Their definitions are provided in the Appendix A.2. Figure 2 shows the specific details.

L1 Perception includes Scene Perception (SP), Target Perception (TP), and Action Perception (AP). SP covers scene type, viewpoint type, and static spatial relation. TP contains target recognition, target counting, and referring spatial localization. AP involves action recognition, action transition, and action counting.

L2 Reasoning contains Complex-environment Reasoning (CER), Spatiotemporal Evolution Reasoning (SER), Spatiotemporal Localization Reasoning (SLR), and Spatiotemporal Consistency Reasoning (SCR). CER covers event reasoning and risk assessment. SER covers spatial relation understanding. SLR covers local-event, single-target, and multi-event spatiotemporal localization. SCR covers spatial relation consistency and target consistency. These seven capability groups also serve as group-level evaluation metrics.

3.2. Dataset Construction

RSVideo-10K is built from eight public sources spanning UAV and satellite videos. To construct this datset, we first decode source videos and identify continuous clips with sufficient visual evidence. Selected clips contain observable targets, actions, scene states, spatial relations, temporal changes, or cross-frame consistency cues. We exclude clips with severe blur, invalid viewpoints, weak temporal continuity, heavy occlusion, or insufficient evidence.

For each task, we design an automated pipeline guided by ground-truth annotations and large-model assistance. Candidate answers are constructed from visually similar categories, neighboring regions, related actions, or easily confused spatial and temporal relations. For cases where the available visual evidence is insufficient for a reliable judgment, we add an "insufficient evidence" answer option. This design avoids forced guessing and evaluates whether models can identify when the available evidence is insufficient for answering the question. The detailed pipeline is provided in Appendix B.1.

3.3. Data Quality and Release

We construct RSVideo-10K through a multi-stage quality control process, including clip screening, question generation, evidence verification, and multi-round review. Three senior domain experts conducted annotation, independent verification, and final adjudication to ensure the quality of videos, questions, answer choices, evidence annotations, and task taxonomy. Qwen-based models were used only for auxiliary consistency checking, while final decisions were made by human annotators. The complete construction procedure is described in Appendix B.1.

All source videos are collected from publicly available resources. We carefully review their licenses and release RSVideo-10K following a license-compliant policy. For redistributable sources, processed clips are released under their corresponding licenses. Source-level license information and release details are provided in Appendix A.1.

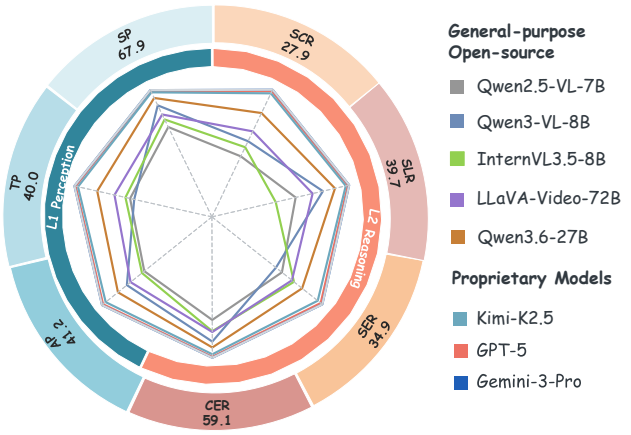


Figure 3. Performance of representative vision-language models on RSVideo-Bench across seven capability groups. The values on the outermost ring represent the results of the proprietary model Gemini-3-Pro on RSVideo-Bench.

4. Findings and Method

We conduct extensive evaluations of existing open-source and proprietary models on RSVideo-Bench. The complete experimental results are reported in Appendix E.. In this section, we first summarize the empirical findings and then introduce RSVideo in detail.

4.1. Benchmark Findings

Figure 3 presents the performance of representative open-source and proprietary vision-language models across seven RSVideo-Bench capability dimensions. The models perform relatively well on SP and CER, but drop clearly on AP, TP, SER, SLR, and SCR, leading to following findings:

Sparse evidence is easily missed. Errors often concentrate on small targets, subtle actions, and short-lived states. The decisive evidence is frequently sparse in both space and time, occupying only a few visual tokens across a limited number of frames. Once these local cues are overlooked, simply increasing model capacity or extending the video context provides limited improvement.

Scene context is necessary but highly redundant. Large remote-sensing scenes contain extensive background regions, such as fields, water bodies, and vegetation, most of which are irrelevant to a given question. However, informative structures, including roads, buildings, region boundaries, and nearby objects, often provide essential spatial cues for reasoning. Therefore, the model needs to preserve useful scene context while avoiding the cost of retaining the entire background at full resolution.

Frame-level recognition is insufficient for spatiotemporal reasoning. Low performance on SER, SLR, and SCR indicates that recognizing individual visual cues is insufficient for coherent video understanding. These tasks require models to associate sparse evidence with the correct target, temporal stage, and spatial context across frames. Missing or misaligned evidence can therefore lead to incorrect reasoning outcomes.

Correct answers do not guarantee grounded decisions. Achieving the correct answer does not necessarily indicate that the model identifies the underlying visual evidence. A model may select the correct option through spurious correlations without associating its prediction with the relevant frames and regions. Reliable video understanding therefore requires not only answer accuracy but also evidence-grounded decision making. These observations motivate the need for explicit supervision that connects model predictions with their supporting spatiotemporal evidence.

4.2. RSVideo

Motivated by the above findings, RSVideo treats evidence as a compact spatiotemporal support set that explains an answer, encoded by temporal positions and spatial cells, e.g., `Evidence: T08-R05/R06; Answer: C`. Training follows a two-stage design. We first perform evidence-aware supervised fine-tuning (E-SFT) on the full visual sequence to establish the evidence format and a reliable policy prior. We then enable sparse visual selection and optimize the token scorer, background allocator, and answer/evidence generation with reinforcement learning. As shown in Figure 4, the RL-stage method contains three parts: spatiotemporal evidence focusing, sparse background compression, and evidence-aware reward optimization.

Spatiotemporal Evidence Focusing. Given a video V and a question Q, we sample T frames, extract patch-level visual features, project them into the language-model embedding space, and flatten all frame tokens into a unified sequence. The question is encoded into a pooled semantic representation:

$$\begin{aligned} \mathbf{H} &= [\mathbf{v}_1; \ldots; \mathbf{v}_L] = \mathrm{Flatten}\left(\mathrm{Proj}(f_{\mathrm{vis}}(V))\right), \\ \mathbf{q} &= \mathrm{Pool}\left(f_{\mathrm{txt}}(Q)\right), \end{aligned} \tag{1}$$

where $\mathbf{H} \in \mathbb{R}^{L \times d}$, $L = T \times N_p$, and N_p is the number of patch tokens per frame.

To rank visual evidence, each token $\mathbf{v}_i$ receives three complementary signals: global saliency s_i^{sal} from visual self-attention, question relevance s_i^{rel} from cross-modal similarity, and temporal change s_i^{chg} from local spatiotemporal inconsistency. After per-video normalization, we fuse them as

$$u_i = \alpha_{\mathrm{sal}} \bar{s}_i^{\mathrm{sal}} + \alpha_{\mathrm{rel}} \bar{s}_i^{\mathrm{rel}} + \alpha_{\mathrm{chg}} \bar{s}_i^{\mathrm{chg}}. \tag{2}$$

Since small remote-sensing targets may be weak at single-token level, we further aggregate responses within coarse

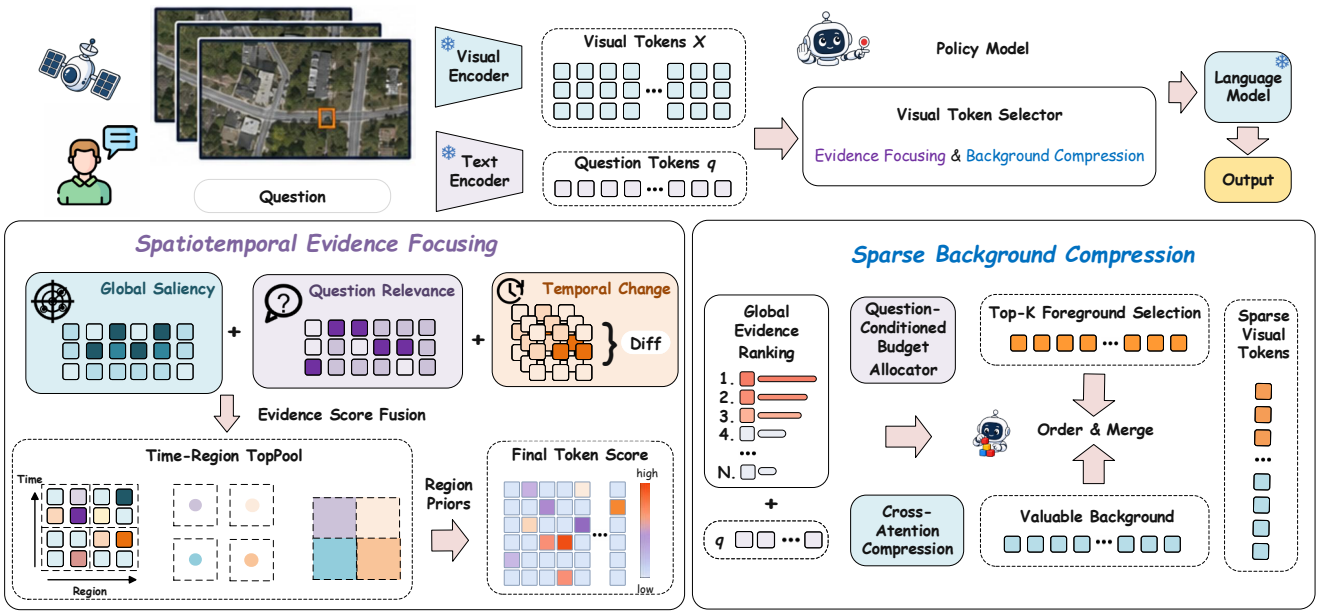


Figure 4. **Overview of RSVideo**. The visual encoder and text encoder first generate visual tokens and question tokens. RSVideo combines global saliency, question relevance, and temporal change for spatiotemporal evidence focusing, integrated with spatiotemporal region priors to obtain final evidence scores for visual tokens. Then RSVideo performs sparse background compression. Based on the evidence scores, RSVideo selects high-value spatiotemporal tokens through global ranking and compresses low-value background tokens into question-related contextual slots. It finally orders and merges the selected foreground tokens and compressed background slots into a sparse visual sequence for the vision-language model to generate the answer.

time–region cells and assign the normalized cell prior back to each token:

$$\ell_i = u_i + \alpha_{\text{cell}} \bar{c}_{\tau_i, g_i}, \tag{3}$$

where (τ_i, g_i) denotes the temporal index and coarse spatial region of token i. The final score ℓ_i therefore combines token-level evidence and local region-level support. The definitions of s_i^{sal}, s_i^{rel}, s_i^{chg}, and $c_{\tau,g}$ are given in Appendix D.1.

Sparse Background Compression and Allocation. Directly dropping all low-score tokens may remove useful context such as roads or buildings. Therefore, RSVideo keeps the strongest evidence tokens and compresses the remaining tokens into a small set of question-conditioned background representations.

Given $\boldsymbol{\ell} = (\ell_1, \ldots, \ell_L)$, we keep a fixed visual budget $B = \max(2, \text{round}(\rho L))$, with $\rho = 0.40$ by default. A lightweight allocator conditioned on $\mathbf{q}$ and the score distribution predicts the background ratio γ. During training, γ is sampled from a predicted Beta distribution to explore different evidence/background trade-offs; during inference, its distribution mean is used. The number of compressed background tokens and retained evidence tokens is

$$M = \text{clip}\left(\text{round}(\gamma B), 1, B-1\right), \qquad K = B - M. \tag{4}$$

The K highest-scoring tokens form $\mathbf{H}_{\text{loc}}$, while the remaining tokens form $\mathbf{H}_{\text{rem}}$, both kept in their original spatiotemporal order.

To preserve compact scene context, we construct M question-conditioned background queries and let them attend to $\mathbf{H}_{\text{rem}}$:

$$\begin{aligned} \mathbf{q}_m^{\text{bg}} &= \mathbf{e}_m + \mathbf{W}_{\text{bg}}\mathbf{q}, \quad m = 1, \ldots, M, \\ \mathbf{Z}_{\text{bg}} &= \text{Attn}\left(\mathbf{Q}_{\text{bg}}, \mathbf{H}_{\text{rem}}, \mathbf{H}_{\text{rem}}\right), \\ \widehat{\mathbf{H}} &= \text{Concat}\left(\mathbf{H}_{\text{loc}}, \mathbf{Z}_{\text{bg}}\right), \end{aligned} \tag{5}$$

where $\mathbf{Q}_{\text{bg}} = [\mathbf{q}_1^{\text{bg}}; \ldots; \mathbf{q}_M^{\text{bg}}]$. The final sequence $\widehat{\mathbf{H}}$ has fixed length $K + M = B$ and is concatenated with the question embeddings before being fed into the language model. The score summary, Beta parameterization, and deterministic inference rule of this allocator are provided in Appendix D.2.

Evidence-Aware Reward and Policy Optimization. After E-SFT, we enable sparse visual selection and optimize the policy with evidence-aware rewards. For each sampled trajectory, the model predicts an answer A and an evidence tag E. The first key reward aligns the predicted evidence with annotated temporal positions and spatial cells. Let $\widehat{\mathcal{T}}(E)$ and $\widehat{\mathcal{G}}(E)$ be the temporal indices and spatial cells parsed from E, and let $\mathcal{T}^*$ and $\mathcal{G}^*$ denote the corresponding

annotations. We compute

$$R_{\mathrm{T}} = \frac{|\widehat{\mathcal{T}}(E) \cap \mathcal{T}^*|}{|\mathcal{T}^*|}, \qquad R_{\mathrm{G}} = \frac{|\widehat{\mathcal{G}}(E) \cap \mathcal{G}^*|}{|\mathcal{G}^*|},$$
$$R_{\mathrm{st}} = \begin{cases} (R_{\mathrm{T}} + R_{\mathrm{G}})/2, & \mathcal{T}^*, \mathcal{G}^* \text{ available}, \\ R_{\mathrm{T}}, & \mathcal{T}^* \text{ only}, \\ R_{\mathrm{G}}, & \mathcal{G}^* \text{ only}. \end{cases} \tag{6}$$

Evidence alignment specifies which spatiotemporal regions should support the answer, but it does not by itself enforce efficient visual routing. We therefore map target annotations to visual-token indices $\mathcal{I}^* \subseteq [L]$. Let $\mathcal{I}_K$ be the Top-K retained indices, $\mathcal{I}_{\mathrm{rem}} = [L] \setminus \mathcal{I}_K$ the residual tokens sent to background compression, and $\mathcal{I}^*_{\mathrm{bg}} = [L] \setminus \mathcal{I}^*$ the non-target tokens. The background compression reward is

$$R_{\mathrm{bg}} = \frac{|\mathcal{I}_{\mathrm{rem}} \cap \mathcal{I}^*_{\mathrm{bg}}|}{|\mathcal{I}^*_{\mathrm{bg}}|} + \frac{|\mathcal{I}_{\mathrm{rem}}|}{L\,M}. \tag{7}$$

The first term rewards routing redundant non-target tokens to the compression branch, while the second favors summarizing more residual tokens with fewer background slots.

The overall return combines answer correctness, evidence alignment, background compression, and a cost penalty for overly broad evidence tags:

$$\mathcal{R} = \lambda_{\mathrm{ans}} R_{\mathrm{ans}} + g_{\mathrm{ans}} \left(\lambda_{\mathrm{st}} R_{\mathrm{st}} + \lambda_{\mathrm{bg}} R_{\mathrm{bg}} - \lambda_{\mathrm{cost}} C_{\mathrm{cost}} \right), \tag{8}$$

where $R_{\mathrm{ans}} = \mathbb{I}[A = A^*]$, and $g_{\mathrm{ans}} = \mathbb{I}[A = A^*]\mathbb{I}[E \text{ is valid}]$ gates the evidence rewards to trajectories with a correct answer and a parseable evidence tag. We optimize the policy using Group Relative Policy Optimization (GRPO) [27]. More details are provided in the Appendix D.5.

5. Experiments

5.1. Experimental Setup

Training follows the two-stage protocol: evidence-aware SFT on the full visual sequence, followed by GRPO with sparse evidence focusing and background compression enabled. All training is conducted on 8× NVIDIA 80 GB A100 GPUs. Unless otherwise specified, we use LoRA fine-tuning with BF16 precision, a maximum sequence length of 2048, gradient checkpointing, per-device batch size 1, gradient accumulation 8, and a cosine learning-rate schedule. The base learning rate is 1×10^{-5}, with warmup ratio 0.03 and weight decay 0.01. LoRA uses rank 32, alpha 64, and dropout 0.05. Complete optimization settings and preprocessing details are provided in Appendix D.7. Accuracy is the proportion of correctly answered questions among all evaluated questions. SP, TP, AP, CER, SER, SLR, and SCR denote the accuracies of the seven capability groups. T-Hit (TH) and R-Hit (RH) measure whether the predicted temporal index T overlaps the annotated key frame or temporal window, and whether the predicted spatial grid index R overlaps the annotated target region, respectively.

Model	Direct	SFT	E-SFT	OG	TG	GSPO	**Ours**
Open-source VLMs: < 7B							
InternVL3-1B	20.13	21.08	22.32	23.08	24.37	24.01	**25.73**
InternVL3.5-1B	20.46	21.55	22.91	23.62	24.18	24.67	**26.23**
InternVL3-2B	21.92	23.10	24.54	25.11	25.96	26.42	**27.87**
InternVL3.5-2B	22.34	23.77	25.16	25.79	26.38	26.91	**28.62**
Qwen3-VL-2B	22.91	24.46	26.08	26.64	27.79	27.38	**29.61**
Qwen2.5-VL-3B	21.68	22.97	24.21	25.02	26.06	25.49	**27.68**
Qwen3-VL-4B	25.46	27.53	29.86	30.47	31.34	31.82	**33.56**
Open-source VLMs: 7–9B							
Qwen2.5-VL-7B	24.88	26.24	28.41	29.02	29.73	30.16	**31.84**
LLaVA-OneVision-7B	23.62	24.72	26.37	27.18	27.84	28.31	**29.89**
InternVL3-8B	26.78	28.68	30.62	31.23	32.57	32.11	**34.36**
InternVL3.5-8B	25.87	27.46	29.78	30.37	31.72	31.16	**33.59**
Qwen3-VL-8B	29.74	31.71	33.68	34.22	35.18	35.61	**37.27**
GLM-4.6V-Flash-9B	30.48	32.05	34.07	34.76	35.38	36.09	**37.83**
Open-source VLMs: 10–30B							
InternVL3-14B	29.12	31.25	33.62	34.31	35.24	35.81	**37.21**
InternVL3.5-14B	28.36	30.52	33.31	33.88	35.27	34.79	**37.37**
Qwen3.6-27B	35.30	35.92	36.78	37.62	37.99	38.47	**40.63**
Open-source VLMs: 30–70B							
Qwen2.5-VL-32B	30.92	32.77	34.82	35.47	36.31	36.78	**38.29**
Qwen3-VL-32B	32.74	34.44	36.75	37.56	38.77	38.31	**40.38**
InternVL3-38B	31.54	33.72	35.83	36.42	37.31	37.78	**39.61**
InternVL3.5-38B	31.08	33.22	35.12	35.81	36.89	36.47	**38.79**
Open-source VLMs: 70–80B							
Qwen2.5-VL-72B	31.36	33.38	35.24	35.82	36.41	37.03	**38.68**
LLaVA-OneVision-72B	28.74	30.21	32.11	32.68	33.95	33.52	**35.47**
LLaVA-Video-72B	29.95	31.72	33.89	34.52	35.11	35.68	**37.42**
InternVL3-78B	33.02	35.06	37.16	37.73	38.66	38.21	**40.36**
Open-source VLMs: > 200B							
Qwen3-VL-235B	33.62	35.67	37.21	37.92	38.58	39.07	**40.49**
InternVL3.5-241B	33.14	35.06	36.98	37.53	38.84	38.36	**40.31**

Table 2. Comparison of different training strategies across various vision-language backbones on RSVideo-Bench. We compare our RSVideo with Direct Inference, supervised fine-tuning (SFT), evidence-aware SFT (E-SFT), and three policy optimization algorithms: outcome-only GRPO (OG) [27], T-GRPO(TG) [8], and GSPO [50]. The best and second-best results for each backbone are highlighted in bold and underline, respectively.

5.2. Comprehensive Evaluation

We conduct a comprehensive evaluation using a diverse set of advanced vision-language backbones spanning parameter scales from 1B to over 200B. For each backbone, we compare our RSVideo with several representative training strategies. All trainable methods train on RSVideo-Instruct and report results on RSVideo-Bench. To ensure a fair comparison, we adopt consistent experimental settings for all methods.

Table 2 reports the evaluation results on RSVideo-Bench. The gains of E-SFT over Direct Inference confirm that RSVideo-Instruct provides effective supervision for remote-sensing video understanding. Building on E-SFT, RL-based approaches further improve performance, while RSVideo ranks first for every evaluated backbone. Specifically, compared with Direct Inference, RSVideo achieves a maximum absolute improvement of 9.01% with InternVL3.5-14B and attains the highest accuracy of 40.63% with Qwen3.6-27B, demonstrating strong generalization across backbone archi-

E-SFT	R_{ans}	R_{st}	R_{bg}	Accuracy ↑	AP ↑	SER ↑	SLR ↑	SCR ↑	TH ↑	RH ↑
✓	✗	✗	✗	36.78	35.96	31.41	36.48	26.67	50.4	47.4
✓	✓	✗	✗	36.90	36.86	32.12	36.74	26.91	51.2	48.3
✓	✓	✓	✗	38.85	39.12	33.90	37.86	27.76	55.4	52.8
✓	✓	✗	✓	39.10	39.96	31.42	37.61	27.48	52.1	53.9
✓	✓	✓	✓	**40.63**	**41.73**	**35.79**	**39.43**	**29.18**	**57.4**	**54.8**

Table 3. Ablation of the evidence-aware rewards.

tectures.

5.3. Ablation Study

To evaluate the contributions of different reinforcement learning rewards, we use Qwen3.6-27B as the backbone and conduct ablation studies under identical experimental settings. Starting from E-SFT, incorporating only the answer-correctness reward (R_{ans}) provides a further but limited improvement, achieving 36.90% accuracy. Adding the spatiotemporal evidence alignment reward (R_{st}) brings the major performance gain, improving accuracy by 1.95 percentage points and increasing TH/RH by 4.2/4.5 points. The improvements are also reflected in AP and SER, which increase from 36.86% and 32.12% to 39.12% and 33.90%, demonstrating the importance of evidence grounding. Adding only the background compression reward (R_{bg}) instead improves accuracy by 2.20 percentage points and RH by 5.6 points, but yields only a 0.9-point increase in TH and a 0.70-point decrease in SER, indicating that it primarily promotes region-level focusing rather than temporally consistent evidence grounding. Combining all rewards achieves the best performance, improving accuracy from 36.90% to 40.63% and TH/RH from 51.2%/48.3% to 57.4%/54.8%. These rewards are complementary: spatiotemporal alignment improves temporal grounding, while background compression enhances regional focus by reducing redundant context.

6. Conclusion

In this paper, we investigate continuous remote-sensing video understanding and reveal that effective reasoning requires accurately identifying and leveraging sparse spatiotemporal evidence. To facilitate systematic evaluation, we introduce RSVideo-10K, including RSVideo-Instruct for training and validation and RSVideo-Bench as a fixed test set. Comprehensive evaluations under a unified five-choice protocol demonstrate that current models often fail to ground their predictions on relevant targets, frames, and scene contexts. Based on these findings, we propose RSVideo, a reinforcement learning framework for evidence-aware spatiotemporal focusing, which adaptively preserves informative evidence tokens, compresses redundant background context, and optimizes evidence-aware rewards under a fixed visual token budget. Extensive experiments demonstrate that RSVideo consistently outperforms existing training methods across diverse vision-language backbones.

References

[1] Anurag Arnab, Ahmet Iscen, Mathilde Caron, Alireza Fathi, and Cordelia Schmid. Temporal chain of thought: Long-video understanding by thinking in frames. In *Advances in Neural Information Processing Systems*, 2025. 3

[2] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025. 17

[3] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025. 1, 17

[4] Mohammadamin Barekatain, Miquel Martí, Hsueh-Fu Shih, Samuel Murray, Kotaro Nakayama, Yutaka Matsuo, and Helmut Prendinger. Okutama-action: An aerial view video dataset for concurrent human action detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 28–35, 2017. 1, 2, 12

[5] Laila Bashmal, Yakoub Bazi, Mohamad M. Al Rahhal, Mohammad Zuair, and Farid Melgani. CapERA: Captioning events in aerial videos. *Remote Sensing*, 15(8):2139, 2023.

[6] Ilker Bozcan and Erdal Kayacan. AU-AIR: A multi-modal unmanned aerial vehicle dataset for low altitude traffic surveillance. In *Proceedings of the IEEE International Conference on Robotics and Automation*, pages 8504–8510, 2020. 1, 2, 12

[7] Yuzeng Chen, Yuqi Tang, Yi Xiao, Qiangqiang Yuan, Yuwei Zhang, Fengqing Liu, Jiang He, and Liangpei Zhang. Satellite video single object tracking: A systematic review and an oriented object tracking benchmark. *ISPRS Journal of Photogrammetry and Remote Sensing*, 210:212–240, 2024. 3, 12

[8] Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, Junfei Wu, Xiaoying Zhang, Benyou Wang, and Xiangyu Yue. Video-R1: Reinforcing video reasoning in MLLMs. *arXiv preprint arXiv:2503.21776*, 2025. 7

[9] Chaoyou Fu, Yuhan Dai, Yongdong Luo, 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*, pages 24108–24118, 2025. 1, 2, 3, 18

[10] Zhe Gao, Shiyu Shen, Taifeng Chai, et al. VideoTIR: Accurate understanding for long videos with efficient tool-integrated reasoning. *arXiv preprint arXiv:2603.25021*, 2026. 3

[11] Jeremy Andrew Irvin, Emily Ruoyu Liu, Joyce Chuyi Chen, Ines Dormoy, Jinyoung Kim, Samar Khanna, Zhuo Zheng, and Stefano Ermon. TEOChat: A large vision-language assistant for temporal earth observation data. In *International Conference on Learning Representations*, 2025. 3

[12] Kartik Kuckreja, Muhammad Sohail Danish, Muzammal Naseer, Abhijit Das, Salman Khan, and Fahad Shahbaz

Khan. GeoChat: Grounded large vision-language model for remote sensing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27831–27840, 2024. 1, 2

[13] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. LLaVA-OneVision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. 1, 17

[14] Kunchang Li, Yali Wang, Yinan He, et al. MVBench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 3, 18

[15] Siyi Li and Dit-Yan Yeung. Visual object tracking for unmanned aerial vehicles: A benchmark and new motion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2017. 12

[16] Xiang Li, Jian Ding, and Mohamed Elhoseiny. Vrsbench: A versatile vision-language benchmark dataset for remote sensing image understanding. In *Advances in Neural Information Processing Systems*, 2024. 1, 2, 3

[17] Yujie Li, Wenjia Xu, Guangzuo Li, Zijian Yu, Zhiwei Wei, Jiuniu Wang, and Mugen Peng. UniRS: Unifying multi-temporal remote sensing tasks through vision language models. *arXiv preprint arXiv:2412.20742*, 2024. 1

[18] Yuanxin Liu, Shicheng Li, Yi Liu, Yuxiang Wang, Shuhuai Ren, Lei Li, Sishuo Chen, Xu Sun, and Lu Hou. TempCompass: Do video LLMs really understand videos? In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 8731–8772, Bangkok, Thailand, 2024. Association for Computational Linguistics. 3

[19] Sylvain Lobry, Diego Marcos, Jesse Murray, and Devis Tuia. RSVQA: Visual question answering for remote sensing data. *IEEE Transactions on Geoscience and Remote Sensing*, 58 (12):8555–8566, 2020. 1, 2, 3

[20] Zhiming Luo, Di Wang, Haonan Guo, Jing Zhang, and Bo Du. VLRS-Bench: A vision-language reasoning benchmark for remote sensing. *arXiv preprint arXiv:2602.07045*, 2026. 3

[21] Martin Q. Ma, Willis Guo, Aditya Agrawal, Ankit Gupta, Paul Pu Liang, Ruslan Salakhutdinov, and Louis-Philippe Morency. Video active perception: Effective inference-time long-form video understanding with vision-language models. *arXiv preprint arXiv:2605.01662*, 2026. 3

[22] Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. EgoSchema: A diagnostic benchmark for very long-form video language understanding. In *Advances in Neural Information Processing Systems*, pages 46212–46244, 2023. 3

[23] Lichao Mou, Yuansheng Hua, Pu Jin, and Xiao Xiang Zhu. ERA: A data set and deep learning benchmark for event recognition in aerial videos. *IEEE Geoscience and Remote Sensing Magazine*, 8(4):125–133, 2020. 1, 2, 12

[24] Shuo Ni, Tong Wang, Jing Zhang, He Chen, Haonan Guo, Ning Zhang, and Bo Du. UHR-Micro: Diagnosing and mitigating the resolution illusion in earth observation VLMs. *arXiv preprint arXiv:2605.12237*, 2026. 1, 2

[25] Qwen Team. Qwen3.6-27B: Flagship-level coding in a 27B dense model. https://qwen.ai/blog?id=qwen3.6-27b, 2026. Accessed July 2026. 17

[26] Alexandre Robicquet, Amir Sadeghian, Alexandre Alahi, and Silvio Savarese. Learning social etiquette: Human trajectory understanding in crowded scenes. In *Proceedings of the European Conference on Computer Vision*, pages 549–565, 2016. 1, 12

[27] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024. 7

[28] Jintao Sun, Hu Zhang, Donglin Di, Gangyi Ding, and Zhedong Zheng. UAVReason: A unified, large-scale benchmark for multimodal aerial scene reasoning and generation. *arXiv preprint arXiv:2604.05377*, 2026. 2, 3

[29] Zhaorui Sun, Yuhang Liu, Haolin Zhu, Yuxuan Gu, Yiming Zou, Zhen Liu, Gui-Song Xia, Bo Du, and Yongchao Xu. RefDrone: A challenging benchmark for referring expression comprehension in drone scenes. *arXiv preprint arXiv:2502.00392*, 2025. 2

[30] Aysim Toker, Lukas Kondmann, Mark Weber, Marvin Eisenberger, Andrés Camero, Jingliang Hu, Ariadna Pregel Hoderlein, Çağlar Şenaras, Timothy Davis, Daniel Cremers, Giovanni Marchisio, Xiao Xiang Zhu, and Laura Leal-Taixé. DynamicEarthNet: Daily multi-spectral satellite dataset for semantic change segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21158–21167, 2022. 1, 2

[31] Di Wang, Jing Zhang, Minqiang Xu, Lin Liu, Dong-Sheng Wang, Erzhong Gao, Chengxi Han, Haonan Guo, Bo Du, Dacheng Tao, and Liangpei Zhang. MTP: Advancing remote sensing foundation model via multitask pretraining. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 17:11632–11654, 2024. 2

[32] Di Wang, Shunyu Liu, Wentao Jiang, Fengxiang Wang, Yi Liu, Xiaolei Qin, Zhiming Luo, Chaoyang Zhou, Haonan Guo, Jing Zhang, Bo Du, Dacheng Tao, and Liangpei Zhang. GeoZero: Incentivizing reasoning from scratch on geospatial scenes. *arXiv preprint arXiv:2511.22645*, 2025. 2

[33] Fengxiang Wang, Hongzhen Wang, Mingshuo Chen, Di Wang, Yulin Wang, Zonghao Guo, Qiang Ma, Long Lan, Wenjing Yang, Jing Zhang, Zhiyuan Liu, and Maosong Sun. XLRS-Bench: Could your multimodal LLMs understand extremely large ultra-high-resolution remote sensing imagery? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. 2, 3

[34] Junjue Wang, Zhuo Zheng, Zihang Chen, Ailong Ma, and Yanfei Zhong. Earthvqa: Towards queryable earth via relational reasoning-based remote sensing visual question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 5481–5489, 2024. 1, 2, 3

[35] Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. InternVL3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025. 17

[36] Zhecheng Wang, Rajanie Prabha, Tianyuan Huang, Jiajun Wu, and Ram Rajagopal. Skyscript: A large and semantically diverse vision-language dataset for remote sensing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 5805–5813, 2024. 1, 2

[37] Ziyang Wang, Shoubin Yu, Elias Stengel-Eskin, Jaehong Yoon, Feng Cheng, Gedas Bertasius, and Mohit Bansal. Videotree: Adaptive tree-based video representation for LLM reasoning on long videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. 3

[38] Weihao Xuan, Junjue Wang, Heli Qi, Zihang Chen, Zhuo Zheng, Yanfei Zhong, Junshi Xia, and Naoto Yokoya. DynamicVL: Benchmarking multimodal large language models for dynamic city understanding. In *Advances in Neural Information Processing Systems*, 2025. 1, 2, 3

[39] Jihan Yang, Shusheng Yang, Anjali W. Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in space: How multimodal large language models see, remember, and recall spaces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10632–10643, 2025. 3

[40] Zeyuan Yang, Delin Chen, Xueyang Yu, Maohao Shen, and Chuang Gan. VCA: Video curious agent for long video understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20168–20179, 2025. 3

[41] Qian Yin, Qingyong Hu, Hao Liu, Feng Zhang, Yingqian Wang, Zaiping Lin, Wei An, and Yulan Guo. Detecting and tracking small and dense moving objects in satellite videos: A benchmark. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–18, 2022. 3, 12

[42] Yufei Yin, Qianke Meng, Minghao Chen, Jiajun Ding, Zhenwei Shao, and Zhou Yu. VideoARM: Agentic reasoning over hierarchical memory for long-form video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026. 3

[43] Tianyu Yu, Zefan Wang, Chongyi Wang, Fuwei Huang, Zhihui He, Tianchi Cai, Weize Chen, Yuxiang Huang, Bokai Xu, Ning Ding, et al. MiniCPM-V 4.5: Cooking efficient MLLMs via architecture, data, and training recipe. *arXiv preprint arXiv:2509.18154*, 2025. 17

[44] Shuai Yuan, Guancong Lin, Lixian Zhang, Runmin Dong, Jinxiao Zhang, Shuang Chen, Juepeng Zheng, Jie Wang, and Haohuan Fu. FUSU: A multi-temporal-source land use change segmentation dataset for fine-grained urban semantic understanding. In *Advances in Neural Information Processing Systems*, 2024. 1, 2

[45] Yang Zhan and Yuan Yuan. Where does it exist from the low-altitude: Spatial aerial video grounding. In *Advances in Neural Information Processing Systems*, 2025. 2

[46] Yang Zhan and Yuan Yuan. UAVBench and UAVIT-1M: Benchmarking and enhancing MLLMs for low-altitude UAV vision-language understanding. *arXiv preprint arXiv:2603.14336*, 2026. 2, 3

[47] Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, Peng Jin, Wenqi Zhang, Fan Wang, Lidong Bing, and Deli Zhao. VideoLLaMA 3: Frontier multimodal foundation models for image and video understanding. *arXiv preprint arXiv:2501.13106*, 2025. 1, 17

[48] Baining Zhao, Jianjie Fang, Zichao Dai, Ziyou Wang, Jirong Zha, Weichen Zhang, Chen Gao, Yue Wang, Jinqiang Cui, Xinlei Chen, and Yong Li. 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)*, pages 32400–32423, 2025. 3, 18

[49] Manqi Zhao, Shengyang Li, Shiyu Xuan, Longxuan Kou, Shuai Gong, and Zhuang Zhou. SatSOT: A benchmark dataset for satellite video single object tracking. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–11, 2022. 1, 3, 12

[50] Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, Rui Men, An Yang, Jingren Zhou, and Junyang Lin. Group sequence policy optimization. *arXiv preprint arXiv:2507.18071*, 2025. 7

[51] Yuanhao Zou, Shengji Jin, Andong Deng, Youpeng Zhao, Jun Wang, and Chen Chen. A.I.R.: Enabling adaptive, iterative, and reasoning-based frame selection for video question answering. In *International Conference on Learning Representations*, 2026. 3

[52] Zhishan Zou, Guoyan Sun, Zhiwei Wei, Jiancheng Pan, Yujie Li, Mugen Peng, and Wenjia Xu. Self in space: Benchmarking self-awareness and spatial cognition in UAV embodied intelligence. *arXiv preprint arXiv:2607.12477*, 2026. 3, 18

Appendix Contents

A. Dataset Construction Details

A.1. Video Sources, Licensing, and Filtering

RSVideo-10K is constructed from eight publicly available remote-sensing video sources spanning UAV, aerial, overhead, and satellite observations. The sources were selected to expose large scene extent, small or crowded targets, camera motion, temporally sparse events, and cross-platform appearance changes. Source membership alone is not an acceptance criterion: every retained segment passes the clip-level audit below.

AU-AIR [6]. AU-AIR records low-altitude UAV traffic scenes with varied viewpoints and visible road users. Its vehicle-rich sequences support target identification, counting, and spatial-relation questions under camera motion.

DTB70 [15]. DTB70 contains UAV tracking sequences with scale variation, occlusion, deformation, and rapid viewpoint change. We use these temporal patterns to construct moving-target localization, trajectory-continuity, and occlusion-recovery tasks.

ERA [23]. ERA provides aerial videos of human activities and events observed from overhead platforms. Its event-centered clips support action recognition, event-state analysis, abnormal-evidence monitoring, and action–event coupling.

OOTB [7]. OOTB contributes satellite video sequences in which targets occupy only a small fraction of the frame and their appearance changes across time. These properties support small-target motion analysis and state recovery across temporal windows.

Okutama-Action [4]. Okutama-Action captures concurrent human activities from UAV viewpoints in outdoor scenes. Its temporally localized action annotations support questions about action transitions, duration, and intervals in which multiple targets remain jointly visible.

SatSOT [49]. SatSOT contains satellite videos for tracking tiny moving targets under scale change, background interference, and intermittent visibility. We use it for target continuity, small-object localization, and recovery after temporary disappearance.

Stanford Drone Dataset [26]. The Stanford Drone Dataset provides overhead trajectories of pedestrians, bicycles, and vehicles in shared outdoor spaces. Its multi-agent scenes support target identification, co-visibility, trajectory interaction, and spatial-relation reasoning.

VISO [41]. VISO provides satellite videos with sparse, small moving objects against wide-area backgrounds. It supports candidate localization, motion reasoning, and negative-evidence questions in which a proposed target or relation is absent from the observed interval.

For every candidate segment, annotators inspect decodability, temporal continuity, observable target/action evidence, and the absence of unusable corruption or watermarks. They then determine whether the segment supports at least one well-posed task in the taxonomy below. We retain clips only when the correct answer can be established from the released visual input and when a concrete temporal window can be identified. Segments with accidental missing frames, ambiguous targets, severe corruption, duplicated content, or answer cues available only outside the released input are rejected. This procedure produces **4,629 audited evidence clips**, covering **17.02 hours** and **1,473,150 decoded frames**.

Group	Task IDs	Task scope
G1 Scene Perception	T1–T3	scene type; viewpoint type; static spatial relation
G2 Target Perception	T4–T6	target recognition; target counting; reference-image target localization
G3 Action Perception	T7–T9	action recognition; action transition; action counting
G4 Complex-Environment Reasoning	T10–T11	event reasoning; risk assessment
G5 Spatiotemporal Evolution Reasoning	T12–T13	global spatiotemporal understanding; long-term understanding
G6 Spatiotemporal Localization Reasoning	T14–T15	static-spatial dynamization; motion-event matching verification
G7 Spatiotemporal Consistency Reasoning	T16–T17	spatial relation consistency; target consistency

Table 4. The 17 tasks grouped into the seven capability groups used by the main-paper evaluation.

An item whose intended answer is *insufficient evidence* is treated differently from a defective clip. Such an item is retained only when insufficiency itself is a well-posed visual conclusion—for example, all five candidate claims require an unobserved continuation—and independent reviewers verify that no released frame resolves the alternatives. A clip that is merely corrupted or accidentally truncated is rejected rather than converted into an insufficiency item.

License-aware release. The public release distributes derived metadata, questions, answer options, split identifiers, and evaluation code while retaining the original providers' terms. By default, it provides source provenance, temporal boundaries, checksums, and reconstruction instructions rather than repackaging third-party pixels. A source video is redistributed only when its verified source manifest explicitly permits that use; the frozen manifest records the corresponding license identifier. The release does not claim ownership of the underlying videos, and

Leaf	Parent task	Leaf capability	Required observable evidence
01	T1 Scene Type	Scene type	Global scene appearance.
02	T2 Viewpoint Type	Viewpoint type	View geometry and perspective.
03	T3 Static Spatial Relation	Spatial relation between static land objects	Spatial relation in one or more frames.
04	T4 Target Recognition	Target identification	Localized target appearance.
05	T5 Target Counting	Target counting	All qualifying targets in the stated region.
06	T6 Reference-Image Target Localization	Reference-image target localization	Reference cue and target region.
07	T7 Action Recognition	Action recognition	Motion/state evidence across frames.
08	T8 Action Transition	Action transition	Before/after action states.
09	T9 Action Counting	Action counting	Event boundaries for each instance.
10	T10 Event Reasoning	Environmental constraint reasoning	Event plus relevant scene context.
11	T10 Event Reasoning	Target state inference	Target/event interaction evidence.
12	T10 Event Reasoning	Main visible activity recognition	Dominant visible activity over the evidence span.
13	T11 Risk Assessment	Disaster evidence monitoring	Hazard cue and affected region/target.
14	T11 Risk Assessment	Abnormal behavior	Contextual norm and observed deviation.
15	T12 Global Spatiotemporal Understanding	Long-term trajectory summary for small moving targets	Target trajectory over a long temporal window.
16	T12 Global Spatiotemporal Understanding	Displacement and speed estimation	At least two temporal positions per moving target.
17	T13 Long-Term Understanding	Temporal ordering	Separate evidence spans for ordered events.
18	T13 Long-Term Understanding	Action-duration comparison	Comparable action intervals or state durations.
19	T13 Long-Term Understanding	Concurrent visibility interval for multiple targets	Overlapping visibility intervals for queried targets.
20	T14 Static-Spatial Dynamization	Single-event spatiotemporal localization	Event interval and local region.
21	T15 Motion-Event Matching Verification	Regional trajectory co-occurrence	Trajectory, region, and event co-occurrence evidence.
22	T16 Spatial Relation Consistency	Relation-constrained candidate localization	Relation checks across the relevant frames.
23	T16 Spatial Relation Consistency	Reasoned counting of small moving targets	Constraint region plus all qualifying targets.
24	T17 Target Consistency	Cross-window constraint satisfiability	Identity attributes across windows.
25	T17 Target Consistency	Target re-identification after occlusion	Pre/post gap identity evidence.
26	T17 Target Consistency	Minimal constraint relaxation	Candidate comparison under explicit constraints.

Table 5. **Operational definitions of the 26 leaves in RSVideo-10K**. Every item is associated with exactly one task ID and one active leaf ID. The required observable evidence specifies the visual or temporal evidence used to validate an item.

users remain responsible for the original providers' terms.

A.2. Task Taxonomy, Operational Leaves, and Question Schema

The taxonomy has two capability levels, seven capability groups, 17 tasks, and 26 operational leaves. L1 *Perception* covers scene, target, and action understanding. L2 *Reasoning* covers complex-environment reasoning, spatiotemporal evolution reasoning, spatiotemporal localization reasoning, and spatiotemporal consistency reasoning. Table 4 summarizes the operational organization used here.

Table 5 enumerates every released operational leaf. The stable schema uses consecutive identifiers 01–26.

Each item is a fixed five-choice multiple-choice question. Its released record contains a video/evidence-clip identifier, the question, five shuffled options, the gold option, capability-level/group/task/leaf labels, and, when applicable, a visual mark and temporal evidence annotation. Questions require visual inspection rather than a source name or memorized fact. Distractors are drawn from nearby categories, confusable actions, reversed temporal order, or scene-incompatible relations. The answer is one canonical option label.

A.3. Split Construction and Data Statistics

Table 6 summarizes the released training, validation, and test partitions and their roles in model development and final evaluation. Splits are created at the evidence-clip level, so all questions tied to the same released segment remain in one partition. We audit source identifiers, near-duplicate clips, and overlapping temporal ranges to avoid cross-split leakage. RSVideo-Instruct contains 8,042 train/validation instances, and RSVideo-Bench contains 2,731 fixed test instances. The test set includes 653 L1 and 2,078 L2 items. All aggregate scores are computed over this fixed item set; they are not unweighted averages over tasks or sources.

B. Data Engine and Annotation Protocol

B.1. Complete Construction and Annotation Pipeline

The construction procedure converts each accepted source video into an auditable evidence record through the following ordered stages:

(1) **Source registration and decoding.** We record each video's provenance, usage terms, and metadata, decode it into candidate frames, and discard material that is unreadable, temporally unstable, or below the required visual-quality threshold.

(2) **Clip screening.** We locate temporal spans with valid

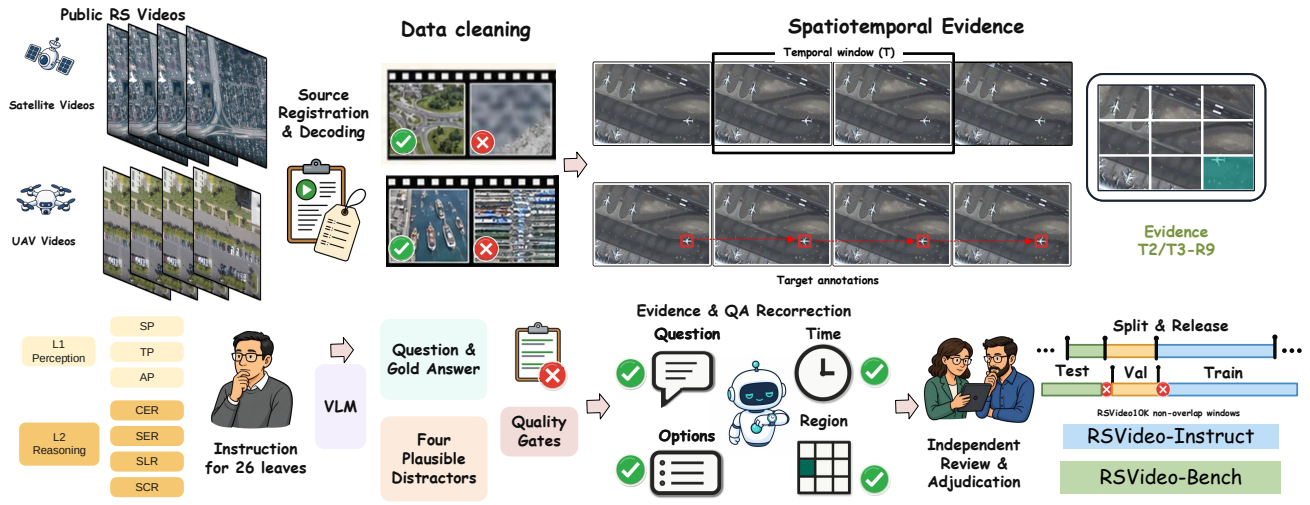


Figure 5. **Complete RSVideo data-construction pipeline.** Public videos first pass source registration, decoding, quality screening, and task/leaf binding before question construction. Each candidate item is then grounded by temporal and, when needed, spatial evidence; paired with one gold answer and four plausible distractors; checked through joint evidence–label correction, option permutation, independent review, and expert adjudication; and finally frozen after clip-level split isolation and release audit. Human verification determines the final labels and evidence fields.

observable content and retain a span only if its released frames support at least one well-posed task.

(3) **Task and leaf binding.** Before question writing, each retained sample is assigned exactly one task ID and one active leaf ID, enforcing a strict correspondence with the predefined taxonomy.

(4) **Spatiotemporal evidence construction.** We annotate the shortest temporal interval sufficient to determine the answer and, for target-centric items, the key spatial region. The interval is expanded only when the surrounding context is necessary to interpret the event or relation.

(5) **Question and answer drafting.** Writers formulate a question around the recorded visible evidence and specify one canonical gold answer, avoiding dependence on source-side context or video content that is not released.

(6) **Candidate construction.** We pair the gold answer with four plausible distractors: categorical items use visually similar classes; localization items use neighboring spatial regions or temporal positions; and action or temporal items use related actions, reversed orders, or confusable state changes. An *insufficient evidence* option is admitted only under the verified rule in Appendix A.1.

(7) **Joint evidence and label correction.** We jointly verify the question, gold answer, task/leaf labels, temporal boundary, spatial mark, and candidate set. If the released video cannot uniquely support a field, the field is corrected when possible; otherwise, the record is discarded.

(8) **Option permutation.** After all semantic fields are frozen, we randomly permute the five answer positions and regenerate the answer key to avoid position bias.

(9) **Independent review and adjudication.** A second expert independently checks answerability, answer uniqueness, evidence sufficiency, and label consistency without seeing the initial annotator's answer. Disagreements enter a third-expert adjudication round, and unresolved records are removed.

(10) **Split and release audit.** We enforce clip-level isolation across training, validation, and test partitions; remove duplicates, near-duplicates, and temporally overlapping material; validate the data schema, stable IDs, option permutations, source manifest, and deterministic evaluation scripts; and finally freeze the release version and checksums.

Split	QA	Role	Access
Train	5,651	supervised training	permitted
Validation	2,391	selection/diagnosis	permitted
Test	2,731	fixed final evaluation	locked
Total	10,773	complete release	—

Table 6. **Released split statistics for RSVideo-10K.** The test partition is RSVideo-Bench; training and validation form RSVideo-Instruct.

Figure 5 summarizes the same pipeline as an end-to-end data engine. The upper flow tracks how public remote-sensing videos are registered, decoded, screened, and bound to task-specific spatiotemporal evidence. The lower flow

records the annotation, verification, adjudication, split isolation, and release-audit stages that convert the evidence into RSVideo-Instruct and RSVideo-Bench records.

The following subsections describe question construction, expert review, adjudication, quality control, and release auditing in greater detail.

B.2. Leaf-guided Question, Answer, and Candidate Construction

Question writers select a task and operational leaf before drafting an item. They bind the item to a visible target, scene property, event, or relation and record the shortest temporal span and, when applicable, spatial region judged sufficient for a decision. This evidence-constrained procedure prevents reliance on hidden source context or an unobserved continuation.

Candidate construction follows the evidence type. Categorical distractors use visually similar categories; localization distractors use neighboring regions or temporal positions; action and temporal distractors use related actions, reversed order, or confusable state changes. An auxiliary large model may flag ambiguous wording, duplicated options, or candidate–evidence mismatches, but it neither assigns the gold answer nor replaces human judgment. The writer resolves every retained issue against the released frames and the recorded task/leaf constraint.

Each retained item has one verified correct option and four plausible distractors. An *insufficient evidence* option is admitted only when the absence of decisive visual support is itself the intended, independently verified answer; it is not a repair for corrupt or accidentally incomplete input. After the gold option, evidence span, and task/leaf label pass verification, option positions are shuffled and the answer key is regenerated.

B.3. Expert Annotation, Multi-round Review, and Adjudication

Three senior domain experts cover initial annotation, independent review, and adjudication. Round 1 freezes a proposed question, five candidates, gold answer, evidence span/region, and task/leaf labels. In Round 2, an independent reviewer inspects the released video without relying on the first annotator's answer and checks answerability, option uniqueness, evidence sufficiency, label consistency, and visual-mark consistency. Any disagreement enters Round 3, in which an adjudicator either establishes one visually supported record or rejects the item. Review stops only when the gold option is unique, the evidence boundary is sufficient, the task/leaf mapping is correct, and all distractors fail for documented visual reasons. Automated VLM outputs remain auxiliary consistency signals and never replace the human decision.

B.4. Quality Control, Label Correction, and Release Audit

Quality control operates at item, clip, split, and release levels. Item-level checks cover decoding, answerability, evidence sufficiency, option uniqueness, task/leaf agreement, and mark consistency. A detected label conflict triggers a joint re-check of question scope, evidence, gold option, task, and leaf: a field is corrected only when the released frames uniquely support the correction; otherwise the item is removed. Clip- and split-level checks reject duplicates, near-duplicates, shared evidence clips, and overlapping temporal ranges across train, validation, and test. Release-level checks validate option permutations, stable identifiers, the source manifest, and deterministic evaluation scripts. Construction traces and adjudication notes remain separate from evaluation prompts so that they cannot act as privileged test information.

C. Evaluation Protocol and Metrics

C.1. Video Input Protocol

Every method receives the same released visual input, question, and five options for an item. The final RSVideo-Bench evaluation uses a fixed video-decoding and frame-sampling procedure, spatial preprocessing policy, prompt template, and answer parser. Methods operate without access to hidden construction records, gold evidence, answer keys, or test-partition annotations. Reported comparisons use the same benchmark examples and one prediction file per run.

C.2. Answer Parsing and Accuracy

The evaluator normalizes a prediction to one of $\{A, B, C, D, E\}$. It accepts an explicit final option label in a structured answer, or an unambiguous option label in the decoded response; responses without a uniquely parsable option are marked incorrect. Let a_n and $\hat{a}_n$ be the gold and parsed predictions for item n. Accuracy is

$$\mathrm{Acc} = \frac{1}{N} \sum_{n=1}^{N} \mathbb{I}[\hat{a}_n = a_n]. \tag{9}$$

The primary score uses all $N = 2{,}731$ items in RSVideo-Bench.

C.3. Evidence Localization Metrics

Methods that emit evidence tags predict a temporal index $\hat{T}$ and one or more spatial grid cells $\hat{R}$. Temporal Hit (T-Hit) is the fraction of items for which $\hat{T}$ overlaps the annotated key frame or temporal window. Region Hit (R-Hit) is the fraction for which $\hat{R}$ overlaps the annotated target region. These are grounding diagnostics, not substitutes for answer accuracy: a response may be correct while relying on incomplete evidence, or localized correctly but choose an incorrect option.

C.4. Score Aggregation

In addition to overall accuracy, we report aggregates by capability level, capability group, task, leaf, and source. Each aggregate is the item-weighted accuracy within its group. Overall accuracy is always calculated directly over all items; it is never obtained by averaging group scores. T-Hit and R-Hit use the same grouping rule when evidence annotations are available. This protocol ensures that all reported views trace back to the same prediction file.

C.5. Validity Audits

The validity audit distinguishes genuine video use from answer priors by comparing text-only, random-single-frame, shuffled-frame, and full-video inputs while retaining the same checkpoint, prompt, option order, answer parser, and test examples. Appendix E.2 reports the resulting capability-wise accuracies from the frozen prediction files.

D. Additional Details of RSVideo

D.1. Evidence Scores

The global saliency score is computed from the visual encoder self-attention. For layer l and head h, let $A^{l,h} \in \mathbb{R}^{L\times L}$ be the attention map normalized over key tokens. We average the attention received by token i over all query tokens, heads, and layers:

$$s_i^{\mathrm{sal}} = \frac{1}{N_l}\sum_{l=1}^{N_l}\frac{1}{L}\sum_{m=1}^{L}\frac{1}{H}\sum_{h=1}^{H} A_{m,i}^{l,h}, \tag{10}$$

where N_l and H are the numbers of visual encoder layers and attention heads.

The question relevance and temporal change scores are

$$\begin{aligned} s_i^{\mathrm{rel}} &= \cos\left(\mathbf{W}_v\mathbf{v}_i,\ \mathbf{W}_q\mathbf{q}\right), \\ s_i^{\mathrm{chg}} &= 1 - \max_{j\in\mathcal{N}(i)} \cos\left(\mathbf{W}_c\mathbf{v}_i,\ \mathbf{W}_c\mathbf{v}_j\right), \end{aligned} \tag{11}$$

where $\mathbf{W}_v$, $\mathbf{W}_q$, and $\mathbf{W}_c$ are learnable projections. The neighborhood $\mathcal{N}(i)$ contains spatially adjacent tokens in the same frame and tokens at identical or nearby patch coordinates in adjacent sampled frames.

Time–Region Cell Prior. Each token has a temporal index $\tau_i \in [T]$ and patch coordinate $\mathbf{p}_i$. A fixed partition function Γ maps $\mathbf{p}_i$ to one of G coarse spatial regions:

$$g_i = \Gamma(\mathbf{p}_i), \qquad \mathcal{C}_{\tau,g} = \{i \in [L] \mid \tau_i = \tau,\ g_i = g\}. \tag{12}$$

For each cell, we estimate evidence strength from its most responsive tokens:

$$\begin{aligned} m_{\tau,g} &= \max\left(1,\ \lceil \eta\,|\mathcal{C}_{\tau,g}|\rceil\right), \\ c_{\tau,g} &= \mathrm{TopMean}_{m_{\tau,g}}\left(\{u_i \mid i \in \mathcal{C}_{\tau,g}\}\right), \end{aligned} \tag{13}$$

where η is the token selection ratio. Appendix E.4 studies the impact of η.

D.2. Budget Allocator

The allocator first summarizes the score distribution as

$$\mathbf{h}_\ell = \left[\mathrm{Mean}(\boldsymbol{\ell}); \mathrm{Std}(\boldsymbol{\ell}); \mathrm{Max}(\boldsymbol{\ell}); \mathrm{TopMean}(\boldsymbol{\ell})\right], \tag{14}$$

where TopMean averages the top 10% scores. It then predicts the Beta parameters for γ:

$$\begin{aligned} (a_\gamma, b_\gamma) &= \mathrm{Softplus}\left(\mathrm{MLP}\left([\mathbf{q}; \mathbf{h}_\ell]\right)\right) + \epsilon, \\ \gamma &\sim \mathrm{Beta}(a_\gamma, b_\gamma). \end{aligned} \tag{15}$$

At inference time, $\gamma = a_\gamma/(a_\gamma + b_\gamma)$, yielding deterministic token allocation.

D.3. Evidence-Aware SFT

In the first training stage, sparsification is disabled and the model receives the complete visual sequence $\mathbf{H}$. Given the target evidence-answer string Y^*, E-SFT optimizes

$$\mathcal{L}_{\mathrm{eSFT}} = -\sum_t \log p_\theta\left(Y_t^* \mid Y_{<t}^*, \mathbf{H}, Q\right), \tag{16}$$

which teaches the model to follow the evidence format before RL.

D.4. Cost Penalty

Let $\mathcal{D}(E)$ be the set of available evidence dimensions (temporal and/or regional). The breadth cost averages only those available dimensions,

$$\begin{aligned} C_{\mathrm{cost}} = \frac{1}{|\mathcal{D}(E)|}\Bigg[&\mathbb{I}[T \in \mathcal{D}(E)]\frac{|\widehat{\mathcal{T}}(E)|}{T} \\ &+ \mathbb{I}[G \in \mathcal{D}(E)]\frac{|\widehat{\mathcal{G}}(E)|}{G}\Bigg]. \end{aligned} \tag{17}$$

When both dimensions are available, this reduces to the one-half average used in the main formulation.

D.5. Group Relative Policy Optimization

For GRPO, we sample a group of responses for each question and compute relative advantages from their evidence-aware returns. Let $\mathcal{R}_k$ be the return of the k-th response in a group. We use

$$\widehat{A}_k = \frac{\mathcal{R}_k - \mathrm{Mean}(\{\mathcal{R}_j\})}{\mathrm{Std}(\{\mathcal{R}_j\}) + \epsilon}, \tag{18}$$

and optimize the clipped policy objective with a KL penalty to the E-SFT policy. This keeps RL updates focused on improving evidence alignment and sparse background routing without drifting from the supervised evidence format.

D.6. Evidence Mapping, Tag Parsing, and Run Records

For a sampled token grid, define the annotation-availability tests

$$c_{\mathrm{T}}(i) = \begin{cases} \mathbb{I}[\tau_i \in \mathcal{T}^*], & \text{if temporal labels exist,} \\ 1, & \text{otherwise,} \end{cases}$$
$$c_{\mathrm{G}}(i) = \begin{cases} \mathbb{I}[g_i \in \mathcal{G}^*], & \text{if regional labels exist,} \\ 1, & \text{otherwise.} \end{cases} \tag{19}$$

The annotation-to-token map is then

$$\mathcal{I}^* = \{i \in [L] \mid c_{\mathrm{T}}(i)c_{\mathrm{G}}(i) = 1\}. \tag{20}$$

Thus, when only one annotation type is present, only its corresponding test restricts the target set.

The canonical output has the form

$$E ::= \texttt{Evidence: } \mathcal{T}_p\text{-}\mathcal{G}_p\texttt{; Answer: } [A\text{–}E],$$
$$\mathcal{T}_p ::= \texttt{T}d(/\texttt{T}d)^*,$$
$$\mathcal{G}_p ::= \texttt{R}d(/\texttt{R}d)^*,$$

where $\mathcal{T}_p$ and $\mathcal{G}_p$ are nonempty lists of temporal and regional indices, respectively, where d denotes a zero-padded nonnegative integer. A deterministic parser extracts the sets, removes duplicates, and checks each index against the sampled-frame count and grid size stored in the run manifest. A trajectory is valid only if it contains exactly one final answer label and every supplied evidence index is in range; malformed or conflicting answer labels are invalid. For items with only temporal or only regional supervision, the unavailable field is omitted and the available field alone determines evidence validity. T-Hit/R-Hit are nonempty-overlap diagnostics, whereas R_{T} and R_{G} above are annotation-coverage fractions.

The four reward weights, overlap conventions, parser version, sampled-frame indices, and grid layout are frozen in the run manifest. The sensitivity report in Appendix E.9 varies these weights around the default configuration while preserving the main-paper optimization protocol.

D.7. Training Configuration, Video Preprocessing, and Reproducibility

Table 7 reports the numerical settings used in the experiments. Training follows E-SFT on the complete visual sequence and then GRPO with sparse evidence focusing and background compression enabled.

Every final run additionally freezes the decoder and version, sampled-frame count and exact indices, temporal stride, spatial resize/crop policy, backbone-native processor and version, token-grid layout, prompt template, evidence/answer parser version, random seed, checkpoint identifier, and prediction-file checksum. The released clip, question, and options are identical across methods; a backbone-native visual processor may differ, but all such differences are recorded with the prediction file.

Setting	Value
Compute	8× NVIDIA A100, 80 GB each
Adaptation / precision	LoRA / BF16
Maximum sequence length	2,048
Memory control	gradient checkpointing
Per-device batch size	1
Gradient accumulation	8
Learning-rate schedule	cosine
Base learning rate	1×10^{-5}
Warmup ratio / weight decay	0.03 / 0.01
LoRA rank / alpha / dropout	32 / 64 / 0.05
Time–region cell pooling ratio	$\eta = 0.10$
Default visual-token budget	$\rho = 0.40$

Table 7. Optimization settings used in the reported experiments.

E. Additional Diagnostic and Robustness Analyses

E.1. Complete Benchmark Results of Existing VLMs

The core results include the 26-backbone in-domain comparison and the progressive reward ablation. This section provides complementary analyses covering benchmark difficulty, validity of video use, transfer evaluation, visual-token budget, inference efficiency, sparse-token integration, RL hyperparameters, reward weights, and qualitative attention patterns. Unless stated otherwise, experiments that use sparse selection use Token Keep 40.0.

To establish the difficulty of RSVideo-Bench before evaluating the proposed training method, we compare general-purpose, remote-sensing-specialized, and proprietary/API VLMs. Table 8 reports all models under the same five-choice benchmark interface, item set, option order, prompt construction, and answer parser; model-native visual processors are retained where required and documented in the run records. Overall accuracy is computed over all 2,731 test instances, and group accuracies follow the seven capability groups defined by the taxonomy. The comparison covers representative MiniCPM, LLaVA, Qwen, VideoLLaMA, and InternVL model families [2, 3, 13, 25, 35, 43, 47].

The strongest evaluated open-source, domain-specialized, and proprietary/API models reach 35.30%, 32.42%, and 40.34% overall accuracy, respectively. Across the evaluated systems, Scene Perception has the highest

Model	Overall ↑	SP ↑	TP ↑	AP ↑	CER ↑	SER ↑	SLR ↑	SCR ↑
Open-source general VLMs								
MiniCPM-V-4.5	22.78	42.13	20.27	23.17	36.36	18.64	21.77	14.93
LLaVA-OneVision-7B	23.62	44.17	21.53	23.87	37.83	19.37	22.54	15.18
Qwen2.5-VL-7B-Instruct	24.88	48.12	23.86	25.38	43.26	22.11	24.23	13.28
VideoLLaMA3-7B	24.94	45.63	22.37	25.14	39.26	20.47	24.23	16.34
InternVL3.5-8B	25.87	52.20	25.21	26.24	47.99	25.91	18.47	15.28
Qwen3-VL-8B-Instruct	29.74	59.75	22.66	31.91	52.35	20.23	32.13	16.60
InternVL3.5-14B	28.36	55.10	27.42	28.31	48.62	23.84	27.46	17.92
LLaVA-OneVision-72B	28.74	53.96	27.88	28.91	47.34	24.16	27.72	18.38
InternVL3-14B	29.12	55.42	28.06	29.28	48.91	24.52	28.34	18.70
LLaVA-Video-72B (Qwen2)	29.95	54.86	28.41	30.52	48.33	25.28	29.14	18.76
GLM-4.6V-Flash	30.48	57.82	28.74	30.15	49.13	24.91	29.82	18.62
Qwen2.5-VL-32B	30.92	57.21	29.34	31.12	50.06	26.18	30.11	19.34
InternVL3.5-38B	31.08	58.04	29.72	31.35	50.42	26.41	30.28	19.46
Qwen2.5-VL-72B	31.36	57.76	29.63	31.57	49.84	26.57	30.43	19.67
InternVL3-38B	31.54	58.66	30.11	31.82	51.06	26.72	30.74	19.82
Kimi-VL-16B-A3B-Instruct	32.09	58.44	30.37	31.63	51.19	26.82	31.46	20.45
Qwen3-VL-32B	32.74	60.42	31.26	33.12	52.34	27.52	32.28	20.62
InternVL3-78B	33.02	60.96	31.84	33.41	52.86	27.83	32.72	20.89
InternVL3.5-241B-A28B-Instruct	33.14	61.62	32.14	33.86	53.28	27.64	33.12	19.81
Qwen3-VL-235B-A22B-Instruct	33.62	61.45	31.84	34.16	52.77	27.86	33.48	21.11
Gemma-4-31B	34.41	62.73	33.57	34.81	54.22	28.79	34.17	21.43
Qwen3.6-27B-Instruct	35.30	63.71	33.44	35.26	54.83	28.49	35.62	22.78
Open-source domain-specialized VLMs								
LLaVA1.5-UAV	28.11	54.83	28.64	26.37	45.92	21.84	24.76	18.63
GeoChat-UAV	28.99	58.21	33.17	27.24	50.46	21.37	24.91	18.42
SIS-Motion-7B	32.42	58.76	30.82	31.69	49.71	28.46	29.57	21.88
Proprietary/API VLMs								
Qwen3-VL-Max	37.62	65.02	37.12	37.64	55.86	31.28	36.96	25.96
Kimi-K2.5	39.27	66.74	39.05	39.86	57.68	33.42	38.74	27.10
GPT-5	39.95	67.58	39.71	40.84	58.72	34.36	39.32	27.55
Gemini-3-Pro	40.34	67.91	40.08	41.23	59.14	34.88	39.72	27.89

Table 8. **Complete RSVideo-Bench results for existing VLMs.** Values are accuracies (%). SP, TP, AP, CER, SER, SLR, and SCR denote Scene Perception, Target Perception, Action Perception, Complex-Environment Reasoning, Spatiotemporal Evolution Reasoning, Spatiotemporal Localization Reasoning, and Spatiotemporal Consistency Reasoning, respectively. Overall and group accuracies are computed from the same frozen prediction file for each model under the protocol in Appendix C.

group accuracy, whereas Spatiotemporal Consistency Reasoning remains substantially lower. This pattern is consistent with the benchmark placing greater demands on temporal and cross-frame relation reasoning than on scene recognition, although the aggregate results do not isolate which visual or architectural factors cause the gap.

E.2. Validity Audit: Dependence on Video Evidence

To test whether performance depends on video evidence rather than only question–option priors or static appearance, we fix a Qwen3.6-27B-Instruct checkpoint fine-tuned on RSVideo-Instruct with full-video inputs and answer supervision, and change only its inference-time input. Table 9 reports the capability-wise results under the fourinference-time input conditions. Text only removes all visual evidence; Random single frame retains static appearance and partial layout; Shuffled frames retains multi-frame content but destroys temporal order; Full video retains both content and order. All four conditions use the same model parameters, questions, option order, prompt, parser, and test set.

Text-only input obtains 23.49%, 3.49 points above the 20.00% chance level, indicating limited but nonzero predictive signal from the question and options. A random frame raises Overall accuracy to 33.17% and slightly exceeds Full video on SP, TP, and CER, indicating that static appearance suffices for many items in those groups. Its AP, SER, SLR, and SCR scores are respectively 8.65, 5.10, 5.83, and 3.46 points below Full video. Shuffled frames reach 35.13%, 1.46 points below ordered Full video, whereas Full video attains the highest Overall, AP, SER, SLR, and SCR scores. Together, these results are consistent with useful contributions from visual content and temporal order, particularly in the four temporally demanding capability groups, while static appearance remains informative for SP, TP, and CER.

E.3. Transfer Evaluation

To test whether the trained checkpoints retain gains beyond the in-domain split, we evaluate each baseline and the corresponding checkpoint trained with RSVideo on RSVideo-Instruct on MVBench, Video-MME, UrbanVideo-Bench, and SIS-Bench [9, 14, 48, 52]. No external benchmark data are used for training or checkpoint selection. Each paired row in Table 10 keeps the backbone and evaluation protocol fixed and changes only whether RSVideo training is applied. Video-MME scores use the without-subtitle setting throughout.

Input Setting	Retained Evidence	Overall ↑	SP ↑	TP ↑	AP ↑	CER ↑	SER ↑	SLR ↑	SCR ↑
Text only	Question and answer options	23.49	28.61	24.27	22.83	25.94	23.47	23.29	20.84
Random single frame	Static appearance and partial layout	33.17	63.47	40.40	27.03	52.11	25.22	29.84	22.20
Shuffled frames	Multi-frame content without correct order	35.13	63.14	40.02	32.52	51.75	28.19	33.10	24.12
Full video	Visual content and temporal order	36.59	62.89	39.69	35.68	51.32	30.32	35.67	25.66

Table 9. Validity audit by inference-time input degradation. All values are accuracies (%), and higher is better (↑). The checkpoint and evaluation protocol remain fixed across rows; only the available visual and temporal evidence changes.

Model / Setting	MVBench ↑	Video-MME ↑	UrbanVideo-Bench ↑	SIS-Bench ↑
InternVL3.5-2B	58.92	67.34	34.76	46.21
InternVL3.5-2B + RSVideo	59.68	67.81	36.19	48.04
Qwen3-VL-4B	62.44	70.86	37.62	50.33
Qwen3-VL-4B + RSVideo	63.31	71.29	39.08	52.11
Qwen2.5-VL-7B	64.78	73.52	39.15	53.46
Qwen2.5-VL-7B + RSVideo	65.42	73.96	40.77	55.02
LLaVA-OV-7B	61.35	72.14	36.83	49.27
LLaVA-OV-7B + RSVideo	62.09	72.48	38.31	50.96
InternVL3-8B	66.91	75.63	41.28	55.19
InternVL3-8B + RSVideo	67.56	76.02	42.91	56.88
GLM-4.6V-Flash-9B	69.47	78.38	43.52	58.64
GLM-4.6V-Flash-9B + RSVideo	70.16	78.74	45.09	60.31
Qwen3.6-27B	75.50	87.70	49.71	68.83
Qwen3.6-27B + RSVideo	76.08	88.04	51.58	70.91
Qwen2.5-VL-72B	72.34	82.16	46.27	63.58
Qwen2.5-VL-72B + RSVideo	72.91	82.49	47.72	65.03
LLaVA-OV-72B	70.82	80.54	43.68	60.42
LLaVA-OV-72B + RSVideo	71.36	80.87	45.01	61.93
InternVL3-78B	73.18	83.41	47.84	65.26
InternVL3-78B + RSVideo	73.79	83.76	49.28	66.71
Qwen3-VL-235B	74.06	85.12	48.63	66.17
Qwen3-VL-235B + RSVideo	74.62	85.38	50.02	67.54

Table 10. Transfer evaluation on external video benchmarks after training on RSVideo-Instruct. Values are overall accuracies (%), where higher is better. Each pair uses the same backbone and evaluation protocol; no external benchmark data are used for training or model selection. OV denotes OneVision.

All eleven evaluated backbones show positive changes on all four external benchmarks after RSVideo training. Averaged over paired backbones, the changes are 0.66, 0.37, 1.52, and 1.64 points on MVBench, Video-MME, UrbanVideo-Bench, and SIS-Bench, respectively. The larger average gains occur on the two remote-sensing-oriented benchmarks, indicating that the RSVideo-trained checkpoints transfer favorably to the tested aerial and satellite video evaluations.

E.4. Time–Region Cell Pooling Ratio

To examine how local evidence is aggregated within each time–region cell, we vary the token selection ratio η while fixing the visual-token budget at $\rho = 0.40$, together with the backbone, training data, checkpoint-selection rule, preprocessing, rollout budget, and evaluator. A smaller η emphasizes the most responsive tokens in each cell, whereas a larger η incorporates broader local context but may dilute sparse evidence with background responses. Table 11 compares $\eta \in \{0.05, 0.10, 0.20\}$ under the same Qwen3.6-27B protocol with the default configuration.

The results show that the cell-level pooling ratio controls the balance between concentrated evidence and surrounding context. Narrow pooling at $\eta = 0.05$ relies on very few highly responsive tokens and is therefore more sensitive to isolated or noisy responses. Broad pooling at $\eta = 0.20$ incorporates more tokens from each cell but can dilute sparse target evidence with background responses. The selected setting, $\eta = 0.10$, achieves the highest Accuracy, TH, and RH, indicating that aggregating the top 10% of tokens provides a suitable balance between evidence concentration and local contextual support.

Pooling Condition	η	Accuracy ↑	TH ↑	RH ↑
Narrow pooling	0.05	39.91	56.6	53.7
Selected default	0.10	40.63	57.4	54.8
Broad pooling	0.20	40.14	56.9	54.2

Table 11. Sensitivity to the time–region cell pooling ratio under the fixed Qwen3.6-27B protocol. η controls the proportion of highest-scoring tokens aggregated within each time–region cell; Accuracy, TH, and RH are percentages, where higher is better.

Budget Condition	ρ	Accuracy ↑	TH ↑	RH ↑
Lower-budget run	0.30	39.72	56.1	53.2
Reported default	0.40	40.63	57.4	54.8
Higher-budget run	0.50	40.18	57.0	54.1

Table 12. Budget sensitivity under the fixed Qwen3.6-27B protocol. ρ denotes the retained visual-token budget ratio; Accuracy, TH, and RH are percentages, where higher is better.

E.5. Visual-token Budget and Background Context

To quantify the accuracy–evidence trade-off, we vary the retained local-evidence budget while fixing the backbone, training data, checkpoint-selection rule, preprocessing, rollout budget, and evaluator. Table 12 compares the selected ratio $\rho = 0.40$ with lower and higher budgets under the same Qwen3.6-27B protocol. All configuration and hyperparameter decisions are made on the RSVideo-Instruct validation split; after the configuration is fixed, RSVideo-Bench is evaluated once.

Among the three tested ratios, $\rho = 0.40$ gives the highest accuracy, TH, and RH, reaching 40.63%, 57.4%, and 54.8%, respectively. Reducing ρ to 0.30 lowers accuracy by

Setting	Memory (GB) ↓	Latency (s) ↓	Acc. ↑
Dense E-SFT	72.6	5.27	36.78
RSVideo	67.1	3.79	40.63

Table 13. Inference efficiency of Qwen3.6-27B on RSVideo-Bench under a fixed single-GPU protocol. Memory denotes peak allocated GPU memory, Latency is the mean end-to-end time per question, and Acc. is test accuracy (%); lower is better for memory and latency, whereas higher is better for accuracy.

0.91 points and reduces both grounding metrics, whereas increasing ρ to 0.50 lowers accuracy by 0.45 points. Thus, $\rho = 0.40$ gives the best observed trade-off among the tested ratios.

E.6. Inference Efficiency

To quantify the inference benefit of sparse visual-token selection, we profile Dense E-SFT and RSVideo on the full RSVideo-Bench test split of 2,731 questions. Both systems use the Qwen3.6-27B backbone and the same NVIDIA A100 80 GB GPU, BF16 inference, batch size 1, deterministic decoding, prompt, and answer parser. Each video is represented by 12 uniformly sampled temporal positions. Peak memory is the maximum allocated GPU memory, and latency is the mean end-to-end time per question over the full split after 20 warm-up examples. Table 13 reports the resulting peak memory, latency, and test accuracy.

On RSVideo-Bench, RSVideo uses 5.5 GB less peak memory (7.6%) and 1.48 s less mean latency per question (28.1%) than dense E-SFT, while improving accuracy by 3.85 points. The smaller memory change relative to the visual-token reduction is expected because model parameters and non-visual states remain resident, whereas the shorter visual sequence directly reduces multimodal prefilling and attention computation. Because the rows represent two complete systems, the accuracy difference is not attributed to sparsification alone.

E.7. Sparse-token Integration

To identify where selection is most effective and how selected evidence is best fused, we compare five integration variants at the same 40% token budget. Table 14 holds the Qwen3.6-27B backbone and training protocol fixed while changing only selection position and fusion strategy.

Post-projector cross-attention achieves the best observed accuracy, 40.63%, exceeding pre-projector pruning, post-projector concatenation, post-projector gated summation, and LLM-layer pruning by 2.21, 1.45, 0.92, and 1.67 points, respectively. These results support the post-projector cross-attention configuration among the tested variants.

Variant	Selection Position	Fusion Strategy	Accuracy ↑
Pre-projector pruning	Before projector	Concatenation	38.42
Post-projector concat	After projector	Concatenation	39.18
Post-projector gated sum	After projector	Gated sum	39.71
Post-projector cross-attention	After projector	Cross-attention	40.63
LLM-layer pruning	LLM layers	Token pruning	38.96

Table 14. Effect of sparse-token integration position and fusion strategy under the fixed Qwen3.6-27B protocol. Higher accuracy is better.

Temperature	Accuracy ↑	TH ↑	RH ↑
0.3	39.46	55.8	53.1
0.5	40.12	56.7	54.0
0.7	40.63	57.4	54.8
1.0	40.21	56.9	54.2
1.5	39.37	55.6	52.9

Table 15. **Effect of Gumbel-Sigmoid temperature.** Accuracy, TH, and RH are reported in %; higher values are better.

Group Size G	Accuracy ↑	TH ↑	RH ↑	Format Valid ↑	Relative Cost ↓
2	39.82	56.3	53.6	98.7	0.50
4	40.63	57.4	54.8	99.1	1.00
6	40.41	57.1	54.4	99.0	1.48
8	40.28	56.8	54.2	98.9	1.96

Table 16. **Effect of GRPO group size G.** Accuracy, TH, RH, and Format Valid are percentages, where higher is better; Relative Cost is normalized to the default group size $G = 4$, and lower is better.

E.8. RL Optimization Sensitivity

To test whether the reported gain depends on a narrow RL configuration, we vary the differentiable selection temperature, GRPO group size, and KL coefficient separately. Each control fixes all other settings. Tables 15, 16, and 17 isolate exploration smoothness, group-relative estimation, and regularization against the evidence-aware SFT reference, respectively.

Temperature 0.7 gives the best observed accuracy, TH, and RH. Temperatures 0.5 and 1.0 remain within 0.51 points of its accuracy, whereas 0.3 and 1.5 are 1.17 and 1.26 points lower. The tested settings therefore favor an intermediate temperature under this protocol.

Group size $G = 4$ gives the highest observed accuracy and format validity at the reference training cost. $G = 2$ halves the reported relative cost but lowers accuracy by 0.81 points, whereas $G = 6$ and $G = 8$ increase relative cost without improving accuracy. Thus, $G = 4$ provides the best observed accuracy–cost trade-off among the tested settings.

Without KL regularization, format validity drops to 96.8% and accuracy reaches 39.74%. A coefficient of 0.03 achieves the best accuracy, 40.63%, with 99.1% valid outputs. Larger coefficients produce slightly higher format

KL Coefficient	Accuracy ↑	Format Valid ↑	Evidence Length ↓
0.00	39.74	96.8	2.84
0.01	40.21	98.5	2.61
0.03	40.63	99.1	2.43
0.05	40.36	99.2	2.31
0.10	39.58	99.4	2.06

Table 17. **Effect of KL regularization against the evidence-aware SFT reference.** Accuracy and Format Valid are reported in %; higher is better (↑). Evidence Length denotes the average number of evidence units, for which lower indicates more compact evidence (↓).

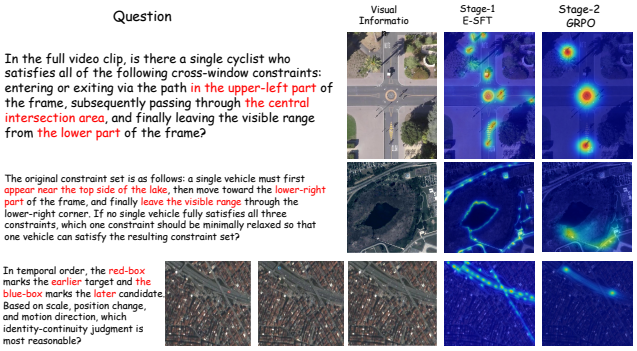


Figure 6. **Qualitative comparison of question-conditioned visual evidence selection before and after reinforcement learning.** From left to right, each row presents the question, the relevant video frame(s), the attention map of the Stage-1 E-SFT model, and that of the Stage-2 GRPO model. The three examples respectively require cross-window trajectory reasoning, minimal relaxation of a spatiotemporal constraint set, and temporal identity-continuity judgment. After GRPO, the attention maps are more concentrated on task-relevant targets, regions, and transition frames.

validity and shorter evidence outputs but lower accuracy, showing that excessive regularization restricts evidence generation.

E.9. Reward-weight Sensitivity

To test whether the full reward depends on a single fragile allocation, we keep every reward component active and vary the relative weights of answer correctness, spatiotemporal evidence alignment, background compression, and evidence-breadth cost. Table 18 uses the same initialization, data, token budget, and RL protocol for every row.

The $0.40/0.40/0.15/0.05$ allocation achieves the best overall accuracy, 40.63%. Assigning more weight to spatiotemporal alignment increases TH from 57.4% to 57.8% but lowers accuracy by 0.35 points, whereas assigning more weight to background compression increases RH from 54.8% to 55.2% but lowers accuracy by 0.56 points. Increasing the evidence-breadth cost weight reduces accuracy to 39.71%. These results show that the default allocation provides the best balance among answer correctness, evidence grounding, and compact evidence selection.

λ_{ans}	λ_{st}	λ_{bg}	λ_{cost}	Accuracy ↑	TH ↑	RH ↑
0.55	0.25	0.15	0.05	39.88	55.9	53.7
0.40	0.40	0.15	0.05	40.63	57.4	54.8
0.30	0.50	0.15	0.05	40.28	57.8	54.6
0.40	0.30	0.25	0.05	40.07	56.5	55.2
0.40	0.35	0.15	0.10	39.71	56.2	53.9

Table 18. **Effect of reward-weight allocation.** λ_{ans}, λ_{st}, λ_{bg}, and λ_{cost} denote answer correctness, spatiotemporal evidence alignment, background compression, and evidence-breadth cost weights; Accuracy, TH, and RH are reported in %, where higher is better.

E.10. Qualitative Evidence Selection Comparison

Figure 6 compares the attention maps produced before and after reinforcement learning. The examples show that GRPO focuses more strongly on question-relevant targets, regions, and transition frames, providing qualitative evidence that the learned policy improves spatiotemporal evidence selection.

F. Leaf-complete Qualitative Examples

We provide one qualitative evidence card for each of the 26 active leaf capabilities in Table 5. Consistent with the taxonomy described above, each leaf specifies a distinct observable evidence pattern rather than a new dataset source or a new evaluation metric. Each card pairs an anonymized frame strip or public mark with the corresponding question, answer options, and visible evidence scope. The examples illustrate what the leaf is designed to test, rather than reporting additional model results.

Leaf 01: Scene Type. Leaf 01 follows the path *L1 Perception* → *G1 Scene Perception* → *T1 Scene Type*. It belongs to perception because the decision rests on directly observable global scene appearance rather than an inferred event or target trajectory.

This leaf tests whether a model can integrate whole-frame context across the clip and select the scene category best supported by that evidence. Local objects may provide useful cues, but the predicted label must characterize the scene as a whole.

As shown in Fig. 7, the example asks which scene category is supported by the clip and contrasts a basketball court, parking lot, baseball field, tennis court, and an evidence-insufficient alternative.

Leaf 02: Viewpoint Type. Leaf 02 follows the path *L1 Perception* → *G1 Scene Perception* → *T2 Viewpoint Type*. It is assigned to scene perception because camera elevation and orientation can be determined from the overall projection geometry of the observed scene.

This leaf assesses whether a model can distinguish viewpoints using perspective, visible object surfaces, and horizon-related

cues. The judgment concerns the imaging configuration rather than the identity or action of any individual target.

As shown in Fig. 8, the example asks the model to choose among top-down aerial, oblique aerial, eye-level ground, low upward-looking, and unsupported viewpoints.

Leaf 03: Spatial Relation Between Static Land Objects. Leaf 03 follows the path *L1 Perception* → *G1 Scene Perception* → *T3 Static Spatial Relation*. It falls under scene perception because the answer is determined by the stable arrangement of land regions relative to a specified reference object.

This leaf evaluates the grounding of directional relations between static scene components. It requires identifying both the reference region and the queried region before comparing their positions.

As shown in Fig. 9, the main road belt is used as the reference, and the model must determine whether the continuous water areas lie mainly in the upper-left, upper-right, lower-right, or left part of the frame, or whether the relation is unsupported by the evidence.

Leaf 04: Target Identification. Leaf 04 follows the path *L1 Perception* → *G2 Target Perception* → *T4 Target Recognition*. It is a target-perception leaf because its output is a semantic category derived from the localized appearance of a visible target.

This leaf tests fine-grained recognition of the queried object while separating target evidence from the surrounding scene context. The model must select a supported category while retaining the option to reject all listed categories when the evidence is insufficient.

As shown in Fig. 10, the example asks which category fits the visible target and contrasts aircraft, person, animal, vessel, and a no-supported-answer option.

Leaf 05: Target Counting. Leaf 05 follows the path *L1 Perception* → *G2 Target Perception* → *T5 Target Counting*. It belongs here because the required operation is to detect and enumerate all visible instances of a specified target class at a given time.

This leaf measures whether a model can avoid missed instances and double counting when the targets are small, crowded, or partially overlapping. The selected count must be supported by the visible evidence at the requested timestamp.

As shown in Fig. 11, the example asks for the number of visible people around 6.7 s and offers counts of 18, 20, 22, and 26, together with an unsupported-choice alternative.

Leaf 06: Reference-image Target Localization. Leaf 06 follows the path *L1 Perception* → *G2 Target Perception* → *T6 Referring Target Localization*. It is categorized as target perception because a referring cue must first be matched to a visible target and then mapped to a spatial region.

This leaf evaluates cue-conditioned localization rather than open-ended target detection. The model must resolve the referred instance and report its position under the supplied grid convention.

As shown in Fig. 12, the example asks where the only visible pedestrian lies in a 3×3 grid around 2.4 s, with upper-right, upper-left, lower-right, center, and unsupported options.

Leaf 07: Action Recognition. Leaf 07 follows the path *L1 Perception* → *G3 Action Perception* → *T7 Action Recognition*. It is placed under action perception because the label is supported by directly visible pose and motion evidence across frames.

This leaf tests recognition of a target's ongoing action without requiring a causal explanation or a longer event chain. Appearance, posture, and short-term motion must be combined to distinguish visually similar actions.

As shown in Fig. 13, the example asks which activity is supported by the video and contrasts walking, standing, running, reading, and an evidence-insufficient alternative.

Leaf 08: Action Transition. Leaf 08 follows the path *L1 Perception* → *G3 Action Perception* → *T8 Action Transition*. It belongs to this task because the answer is defined by a change between two directly observed action states.

This leaf assesses whether a model can segment the clip around an action transition and identify the state on each side of it. Correct recognition of both temporal segments is necessary, as recognizing either state alone is insufficient.

As shown in Fig. 14, the example tracks a red-box person and contrasts remaining lying down throughout, switching from standing to walking, switching from lying down to sitting, switching from sitting to lying down, and an unsupported alternative.

Leaf 09: Action Counting. Leaf 09 follows the path *L1 Perception* → *G3 Action Perception* → *T9 Action Counting*. It is assigned to action perception because it counts boundaries between observable action states rather than counting target instances.

This leaf tests repeated temporal segmentation of a marked target's behavior. The model must distinguish genuine action-state changes from the continued execution of the same action.

As shown in Fig. 15, the example asks how many times the red-box person's action state changes during the clip, with candidate counts ranging from one to five.

Leaf 10: Environmental Constraint Reasoning. Leaf 10 follows the path *L2 Reasoning* → *G4 Complex-Environment Reasoning* → *T10 Event Reasoning*. It belongs to reasoning because the model must relate a target's activity to the surrounding scene structure and infer which environmental constraint best explains the observed motion.

This leaf tests event interpretation under physical layout constraints, including boundaries, passable corridors, curves, intersections, and blockages. The relevant evidence therefore combines target behavior with persistent environmental context.

As shown in Fig. 16, the example asks what spatial constraint governs the observed activity and contrasts a limiting boundary, a turning guide, a road or corridor axis, an occupied passable area, and insufficient evidence.

Leaf 11: Target State Inference. Leaf 11 follows the path *L2 Reasoning* → *G4 Complex-Environment Reasoning* → *T10 Event Reasoning*. It is grouped with event reasoning because the target's state or role cannot be determined from appearance alone and must instead be inferred from its interaction with the surrounding event.

This leaf evaluates contextual inference about why a target appears at a marked location or what event-related state it occupies. The model must compare plausible explanations and reject those inconsistent with the visible sequence.

As shown in Fig. 17, the example asks why a red vehicle appears at the marked location and contrasts prior parking, an opposing race car, rescue arrival, accident-related loss of control, and insufficient evidence.

Leaf 12: Main Visible Activity Recognition. Leaf 12 follows the path *L2 Reasoning → G4 Complex-Environment Reasoning → T10 Event Reasoning*. It is an event-reasoning leaf because the answer summarizes the dominant activity formed by multiple targets and the scene context over the clip, rather than labeling an isolated object or action.

This leaf tests whether a model can aggregate distributed visual evidence into the principal activity represented by the sequence. Incidental objects and background structures should not outweigh the sustained activity pattern.

As shown in Fig. 18, the example asks which category best describes the clip and contrasts a construction site, chase game, campus scene, running race, and an unsupported category.

Leaf 13: Disaster Evidence Monitoring. Leaf 13 follows the path *L2 Reasoning → G4 Complex-Environment Reasoning → T11 Risk Assessment*. It belongs to risk assessment because visible environmental changes or damage cues must be interpreted as evidence of a potential hazard category.

This leaf evaluates the recognition and monitoring of disaster-related evidence while separating visually confusable risk types. The judgment is restricted to cues present in the clip and does not require unsupported claims about the cause or severity of the event.

As shown in Fig. 19, the example asks which environmental evidence category fits the clip and compares water-related, combustion-related, slope-failure, damage-related, and structural-site-change evidence.

Leaf 14: Abnormal Behavior. Leaf 14 follows the path *L2 Reasoning → G4 Complex-Environment Reasoning → T11 Risk Assessment*. It is assigned to risk assessment because the model must decide whether an observed interaction constitutes dangerous behavior and support that judgment with temporal evidence.

This leaf tests the contextual detection of abnormal actions together with approximate event timing. It requires distinguishing the absence of danger from different hazardous interactions occurring at competing timestamps.

As shown in Fig. 20, the example asks whether pushing or pulling occurs around 3 s or 7 s, while also allowing the conclusion that no dangerous action occurs.

Leaf 15: Long-term Trajectory Summary for Small Moving Targets. Leaf 15 follows the path *L2 Reasoning → G5 Spatiotemporal Evolution Reasoning → T12 Global Spatiotemporal Understanding*. It belongs to global spatiotemporal understanding because the trajectory must be integrated over the full clip despite the small size of the moving target.

This leaf tests whether a model can maintain a target track across a long temporal window and compress it into a start-to-end directional summary. Single-frame position cues are insufficient for the required description.

As shown in Fig. 21, the example asks which option best summarizes the small aircraft's long-term trajectory and contrasts motion from the upper-left to the lower-left, from the frame center to the lower-left, from the frame center to the upper-left, and from the lower-left to the lower-right, together with the possibility that no moving aircraft is visible.

Leaf 16: Displacement and Speed Estimation. Leaf 16 follows the path *L2 Reasoning → G5 Spatiotemporal Evolution Reasoning → T12 Global Spatiotemporal Understanding*. It is placed here because the overall movement direction and displacement magnitude must be derived by comparing the target's positions across the observed sequence.

This leaf assesses quantitative motion reasoning at a coarse, evidence-supported resolution. The model must jointly estimate direction and displacement range rather than identify movement in only one frame.

As shown in Fig. 22, the example asks for the aircraft's overall movement direction and displacement magnitude and contrasts upper-right motion over 0–20 pixels, upper-left motion over 60–120 pixels, lower-right motion over 20–60 pixels, lower-right motion over 120–200 pixels, and the possibility that no moving aircraft is visible.

Leaf 17: Temporal Ordering. Leaf 17 follows the path *L2 Reasoning → G5 Spatiotemporal Evolution Reasoning → T13 Local-Temporal Understanding*. It belongs to local-temporal understanding because the required relation is the order of actions across specified neighboring intervals.

This leaf tests whether a model can preserve event chronology while recognizing the action performed in each segment. An answer containing the correct set of actions but presenting them in the wrong order should therefore be rejected.

As shown in Fig. 23, the example orders the red-box target's actions over 0–3 s, 3–5 s, and 5–8 s using combinations of walking, shaking hands, standing, and running.

Leaf 18: Action-duration Comparison. Leaf 18 follows the path *L2 Reasoning → G5 Spatiotemporal Evolution Reasoning → T13 Local-Temporal Understanding*. It is assigned to this task because the answer depends on accumulating and comparing the durations of different action intervals for the same target.

This leaf evaluates temporal measurement beyond action recognition or transition detection. Repeated intervals of the same activity must be combined before selecting the activity with the longest total duration.

As shown in Fig. 24, the example asks which activity occupies the longest total duration for the red-box target, contrasting running, hugging, walking, pushing or pulling, and an unsupported alternative.

Leaf 19: Concurrent Visibility Interval for Multiple Targets. Leaf 19 follows the path *L2 Reasoning → G5 Spatiotem-*

poral Evolution Reasoning → T13 Local-Temporal Understanding. It belongs here because the answer is obtained by intersecting the visibility intervals of multiple marked targets.

This leaf tests synchronized temporal tracking rather than independent target detection. The model must identify when all queried targets are visible simultaneously and select the closest supported interval.

As shown in Fig. 25, the example asks when the red- and blue-marked targets are both visible and compares the intervals 6–12 s, 2–8 s, 0–11 s, 0–6 s, and 10–16 s.

Leaf 20: Single-event Spatiotemporal Localization. Leaf 20 follows the path *L2 Reasoning → G6 Spatiotemporal Localization Reasoning → T14 Single-Event Spatiotemporal Localization*. It is assigned to this task because one event must be detected and localized to the temporal interval in which it occurs.

This leaf tests event-boundary identification for a specified target and event type. The selected interval must closely cover the observed event rather than merely contain the target somewhere in the clip.

As shown in Fig. 26, the example asks when a visible white car disappears and offers 2–3 s, 3–5 s, 5–7 s, and 7–9 s intervals, together with a no-match option.

Leaf 21: Regional Trajectory Co-occurrence. Leaf 21 follows the path *L2 Reasoning → G6 Spatiotemporal Localization Reasoning → T15 Multi-Event Spatiotemporal Localization*. It belongs to multi-event localization because the answer jointly constrains multiple target trajectories by time and region.

This leaf evaluates whether a model can verify co-occurrence only when the queried identities, spatial region, and temporal window all agree. Observing both targets somewhere in the clip is not sufficient.

As shown in Fig. 27, the example asks whether the red-box and blue-box targets are simultaneously visible in the highlighted lower-center region around 6 s and contrasts joint, single-target, and nonconcurrent alternatives.

Leaf 22: Relation-constrained Candidate Localization. Leaf 22 follows the path *L2 Reasoning → G7 Spatiotemporal Consistency Reasoning → T16 Spatial Relation Consistency*. It is placed under spatial relation consistency because a candidate is valid only if its attributes and its relation to the referenced scene region are jointly satisfied.

This leaf tests localization after relational and semantic filtering. The model must first identify candidates that satisfy the stated description and then map the surviving candidate to the requested grid area.

As shown in Fig. 28, the example asks where a local standby vehicle capable of carrying more than 50 people is located on the runway and contrasts the upper-left, upper-center, middle-left, and center regions of a 3×3 grid, together with an unsupported alternative.

Leaf 23: Reasoned Counting of Small Moving Targets. Leaf 23 follows the path *L2 Reasoning → G7 Spatiotemporal Consistency Reasoning → T16 Spatial Relation Consistency*. It belongs to this task because targets contribute to the count only when their positions satisfy a stated spatial-context relation.

This leaf evaluates constrained counting of small targets rather than unfiltered enumeration. The model must apply the spatial predicate consistently to every candidate before producing the total.

As shown in Fig. 29, the example asks how many people are located in areas shadowed by trees or buildings and contrasts counts of five through eight with an unsupported-count alternative.

Leaf 24: Cross-window Constraint Satisfiability. Leaf 24 follows the path *L2 Reasoning → G7 Spatiotemporal Consistency Reasoning → T17 Target Consistency*. It is assigned to target consistency because observations from separate temporal windows must be attributable to one continuous target before the complete constraint set can be satisfied.

This leaf tests whether identity and trajectory evidence remain jointly consistent across multiple temporal windows. Partial satisfaction by different targets must not be mistaken for a single valid track.

As shown in Fig. 30, the example asks whether one cyclist can be linked from the upper-left path through the central intersection to a lower-frame exit and contrasts full continuity with several partial or disconnected cases.

Leaf 25: Target Re-identification after Occlusion. Leaf 25 follows the path *L2 Reasoning → G7 Spatiotemporal Consistency Reasoning → T17 Target Consistency*. It belongs here because an earlier observation and a later candidate must be judged as the same or different target across an interruption in visibility.

This leaf evaluates re-identification using motion-compatible and appearance-compatible evidence around an occlusion or temporal gap. The judgment should reflect plausible scale, displacement, and movement-direction continuity rather than relying on the box colors alone.

As shown in Fig. 31, the example marks the earlier target in red and the later candidate in blue and asks which identity-continuity judgment is most consistent with their scale, position change, and motion direction.

Leaf 26: Minimal Constraint Relaxation. Leaf 26 follows the path *L2 Reasoning → G7 Spatiotemporal Consistency Reasoning → T17 Target Consistency*. It is a target-consistency leaf because the model must preserve a single-target interpretation while determining which one of several cross-time constraints prevents the complete constraint set from being satisfied.

This leaf tests the minimal relaxation of an unsatisfied constraint set. The preferred answer changes only the constraint necessary to obtain a visually supported trajectory while retaining as much of the original specification as possible.

As shown in Fig. 32, the example constrains one vehicle to appear near the top side of a lake, move toward the lower-right, and exit at the lower-right corner, and then asks whether to relax the starting-position, motion-direction, exit-location, or same-vehicle constraint, or make no relaxation.

Classification:

L1 Perception / G1 Scene Perception / T1 Scene Type / 01 Scene type

Question:

Based on the visible spatiotemporal evidence, which scene category is best supported by this clip?

Options:

A. Basketball court

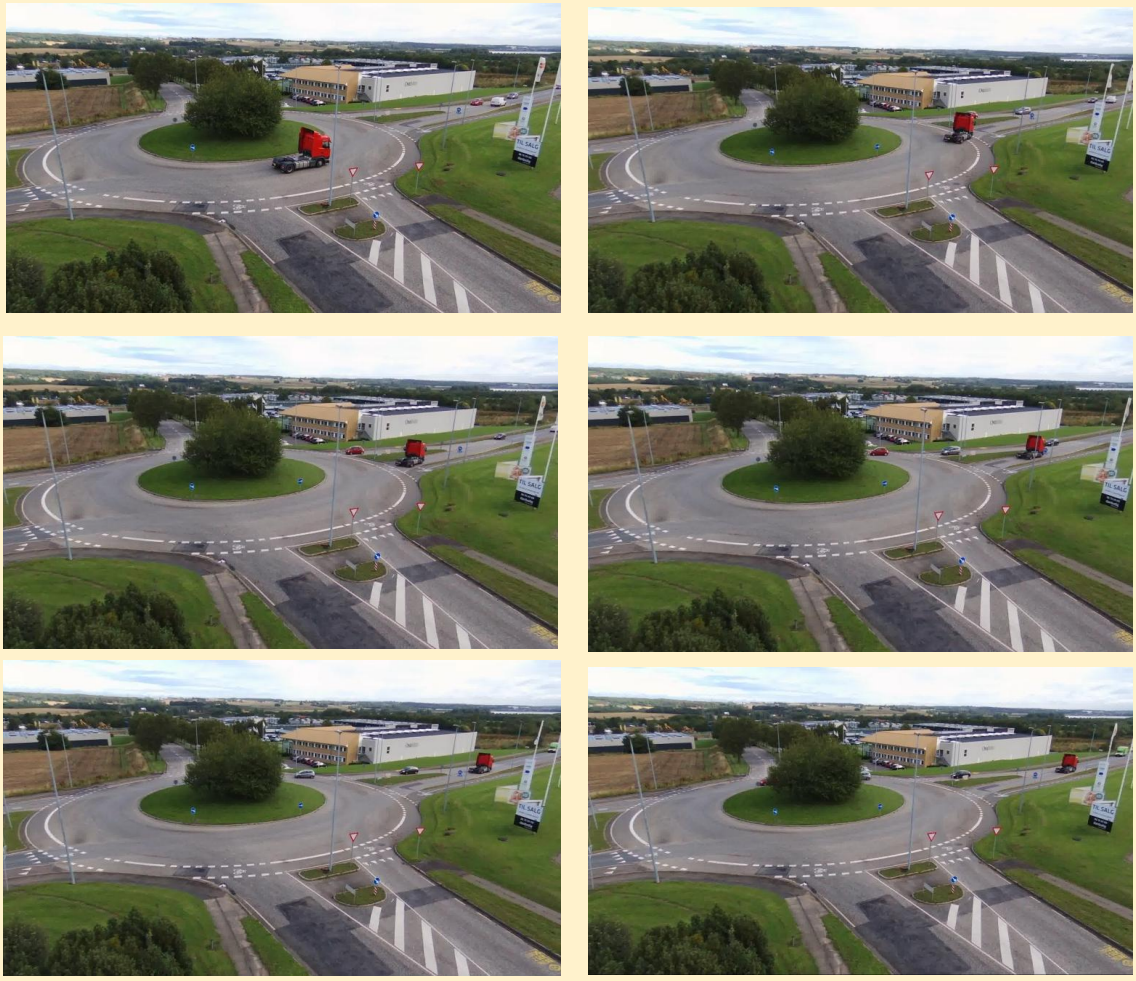
B. None of the listed scene categories is supported by the visible evidence

C. Parking lot

D. Baseball field

E. Tennis court

Figure 7. Qualitative example for Leaf 01, scene type.



Classification:

L1 Perception / Scene Perception / T2 Viewpoint Type / 02 Viewpoint type

Question:

Which camera viewpoint best matches this clip?

Options:

A. Top-down aerial view, with objects mostly seen from above

B. Oblique aerial view, with both object tops and side surfaces visible ✅

C. Eye-level ground view, with the horizon or street-level perspective visible

D. Low upward-looking view, with the camera looking up at objects from below

E. No listed alternative is warranted by the visible evidence

Figure 8. Qualitative example for Leaf 02, viewpoint type.

Classification:

L1 Perception / Scene Perception / T3 Static Spatial Relation / 03 Spatial relation between static land objects

Question:

Using the main road belt as the reference, in which direction are continuous water areas mainly distributed?

Options:

A. None of the listed options is supported by the visible evidence

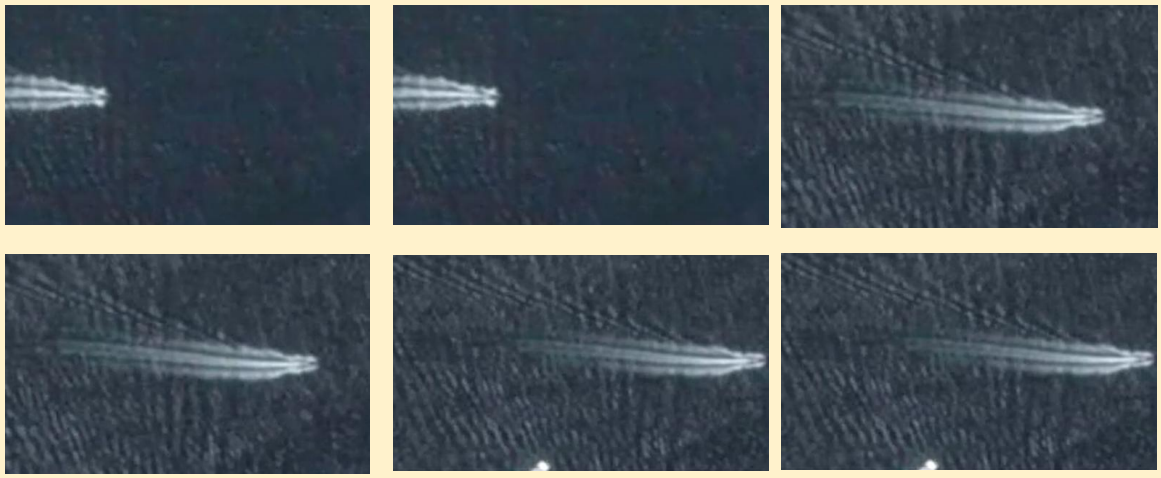
B. Upper-left frame region

C. Upper-right frame region

D. Lower-right frame region ✅

E. Left frame region

Figure 9. Qualitative example for Leaf 03, spatial relation between static land objects.



Classification:

L1 Perception / Target Perception / T4 Target Recognition / 04 Target identification

Question:

Which category best fits the visible target?

Options:

A. The visual evidence supports category aircraft

B. The visible target is an animal

C. The visual evidence supports category person

D. The visible target is a boat or vessel

E. No provided answer is supported by the visible clip.

Figure 10. Qualitative example for Leaf 04, target identification.

Classification:

L1 Perception / Target Perception / T5 Target Counting / 05 Target counting identification

Question:

Around 6.7 s in the clip, which count is most reasonable for the visible people?

Options:

A. The visible clip does not support any provided choice.

B. 18 occurrences are supported

C. 20 occurrences are supported

D. The evidence-backed count is 22

E. The evidence-backed count is 26

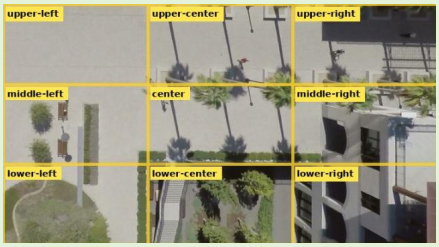
Figure 11. Qualitative example for Leaf 05, target counting.

Classification:

L1 Perception / Target Perception / T6 Referring Target Localization / 06 Reference-image target localization

Question:

At around 2.4 s, using the 3x3 visual grid overlay, which grid region does the only visible pedestrian mainly fall in?



Options:

A. Upper-right frame region

B. Upper-left frame region

C. Lower-right frame region

D. Center frame region

E. None of the listed options is supported by the visible evidence.

Figure 12. Qualitative example for Leaf 06, reference-image target localization.

Classification:

L1 Perception / Action Perception / T7 Action Recognition / 07 Action recognition

Question:

Based on the visible spatiotemporal evidence, which scene category is best supported by this video?

Options:

A. Walking

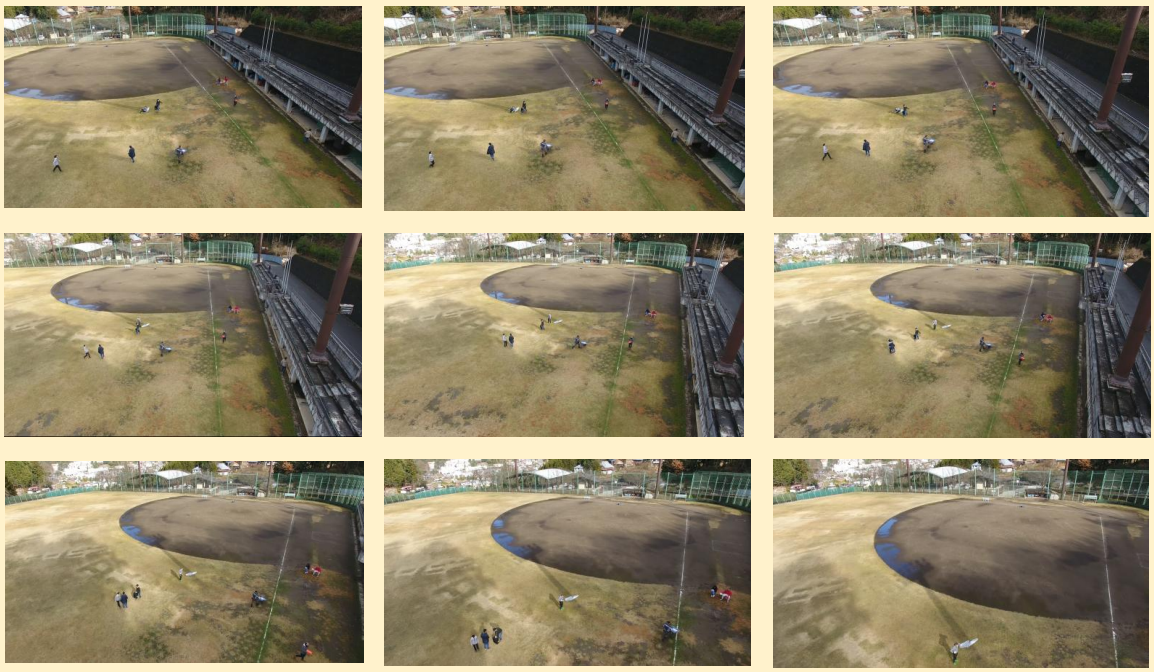
B. Standing

C. Running

D. Reading

E. None of the listed options is supported by the visible evidence.

Figure 13. Qualitative example for Leaf 07, action recognition.



Classification:

L1 Perception / Action Perception / T8 Action Transition / 08 Action transition

Question:

What are the red-box person's actions before and after the action change?

Options:

A. The person remains lying down throughout

B. Switches from standing to walking

C. Switches from lying down to sitting or seated

D. No listed choice is supported after checking the visual evidence carefully.

E. Switches from sitting to lying down

Figure 14. Qualitative example for Leaf 08, action transition.

Classification:

L1 Perception / Action Perception / T9 Action Counting / 09 Action counting

Question:

From the video evidence, how many times does the red-box person's action state change in the clip?

Options:

A. 3 times

B. 2 times

C. 4 times

D. 1 time

E. 5 times

Figure 15. Qualitative example for Leaf 09, action counting.

Classification:

L2 Reasoning / Complex-environment Reasoning / T10 Event Reasoning / 10 Environmental constraint reasoning

Question:

In the full clip, what main constraint does the spatial structure impose on the target activity?

Options:

A. None of the listed options is supported by the visible evidence.

B. A boundary limits the active area

C. A curve or intersection guides turning

D. A road or corridor axis guides the main movement direction

E. Part of the passable area is occupied or blocked

Figure 16. Qualitative example for Leaf 10, environmental constraint reasoning.

Classification:

L2 Reasoning / Complex-environment Reasoning / T10 Event Reasoning / 11 Target state inference

Question:

Which of the following is the most likely reason that the red vehicle appeared at this marked location?

Options:

A. The red vehicle had been parked there all along.

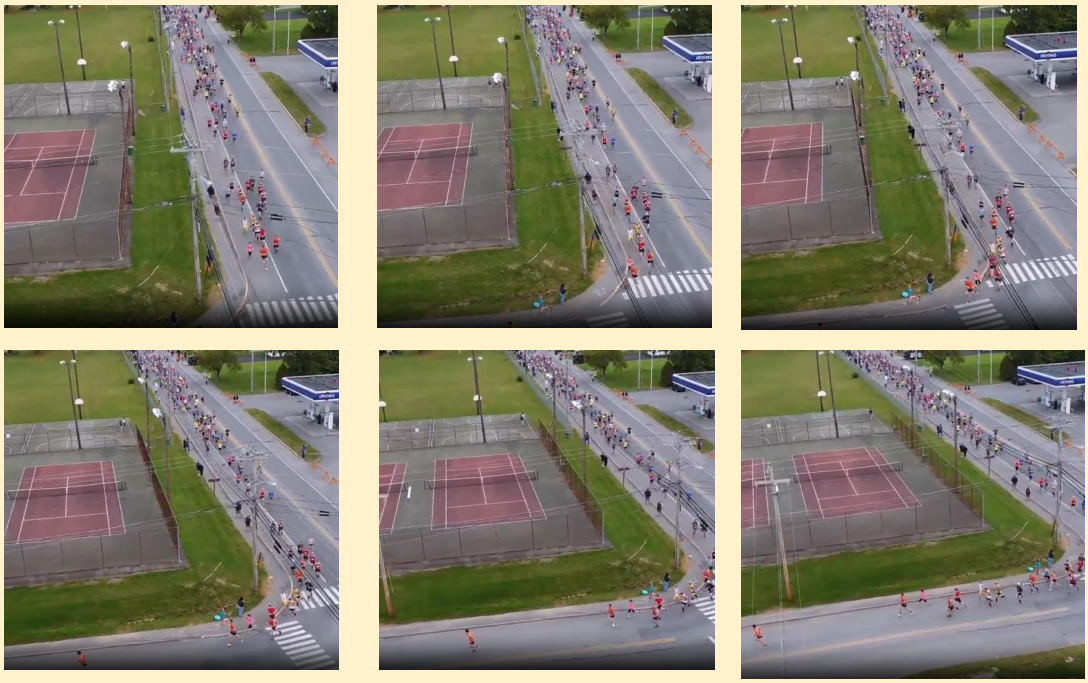
B. The red vehicle was a race car approaching from the opposite direction.

C. The red vehicle was a rescue vehicle arriving to provide assistance.

D. The red vehicle lost control due to a traffic accident involving the black vehicle.

E. There is insufficient evidence to determine the reason.

Figure 17. Qualitative example for Leaf 11, target state inference.



Classification:

L2 Reasoning / Complex-environment Reasoning / T10 Event Reasoning / 12 Main visible activity recognition

Question:

Based on the visible spatiotemporal evidence, which scene category is best supported by this clip?

Options:

A. Construction site

B. Chase Game

C. Campus scene

D. None of the listed scene categories is supported by the visible evidence

E. Running race

Figure 18. Qualitative example for Leaf 12, main visible activity recognition.

Classification:

L2 Reasoning / Complex-environment Reasoning / T11 Risk Assessment / 13 Disaster evidence monitoring

Question:

Based on the visible clip, which environmental evidence category fits the clip best?

Options:

A. Water-related evidence such as flooding or runoff is visible.

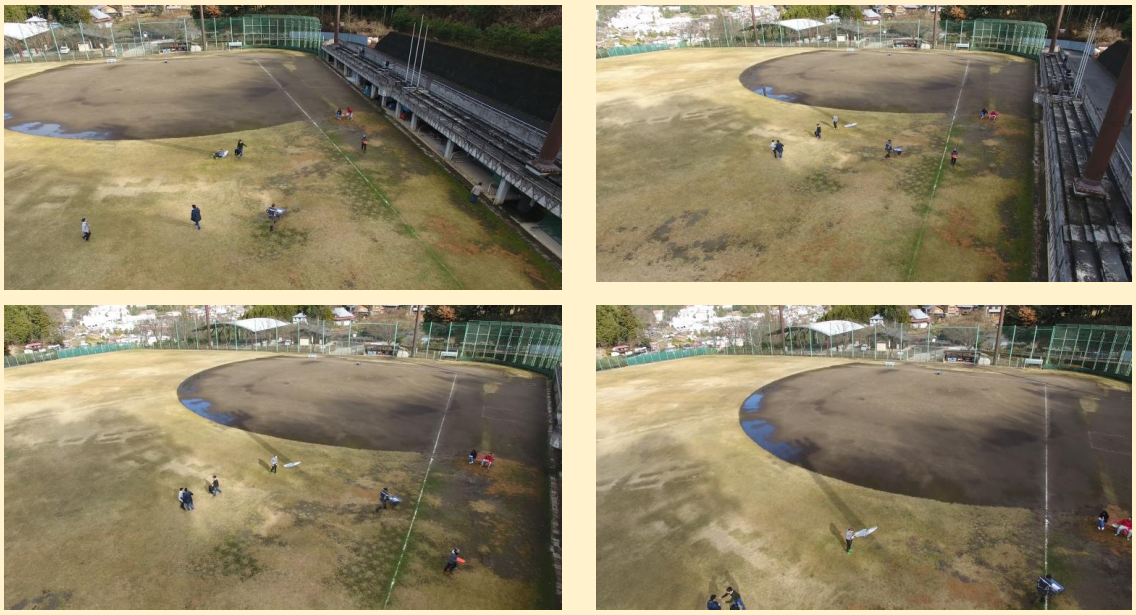
B. Combustion-related evidence such as fire or smoke is visible.

C. Slope-failure evidence such as a landslide or mudslide is visible.

D. Damage-related evidence such as debris, collapse, or collision impact is visible.

E. A structural-site change such as construction or building damage is visible..

Figure 19. Qualitative example for Leaf 13, disaster evidence monitoring.



Classification:

L2 Reasoning / Complex-environment Reasoning / T11 Risk Assessment / 14 Abnormal behavior

Question:

Does any dangerous action occur in this video clip? If so, indicate at approximately which second it occurs.

Options:

A. No dangerous action occurs in the video clip.

B. A dangerous action occurs at around the 3-second mark: pulling a person.

C. A dangerous action occurs at around the 3-second mark: pushing a person.

D. A dangerous action occurs at around the 7-second mark: pulling a person.

E. A dangerous action occurs at around the 7-second mark: pushing a person.

Figure 20. Qualitative example for Leaf 14, abnormal behavior.

Classification:

L2 Reasoning / Spatiotemporal Evolution Reasoning / T12 Global Spatiotemporal Understanding / 15 Long-term trajectory summary for small moving targets

Question:

Based on the full clip, which option best summarizes the small aircraft's long-term trajectory from start to end??

Options:

A. No moving aircraft target is visible in the video.

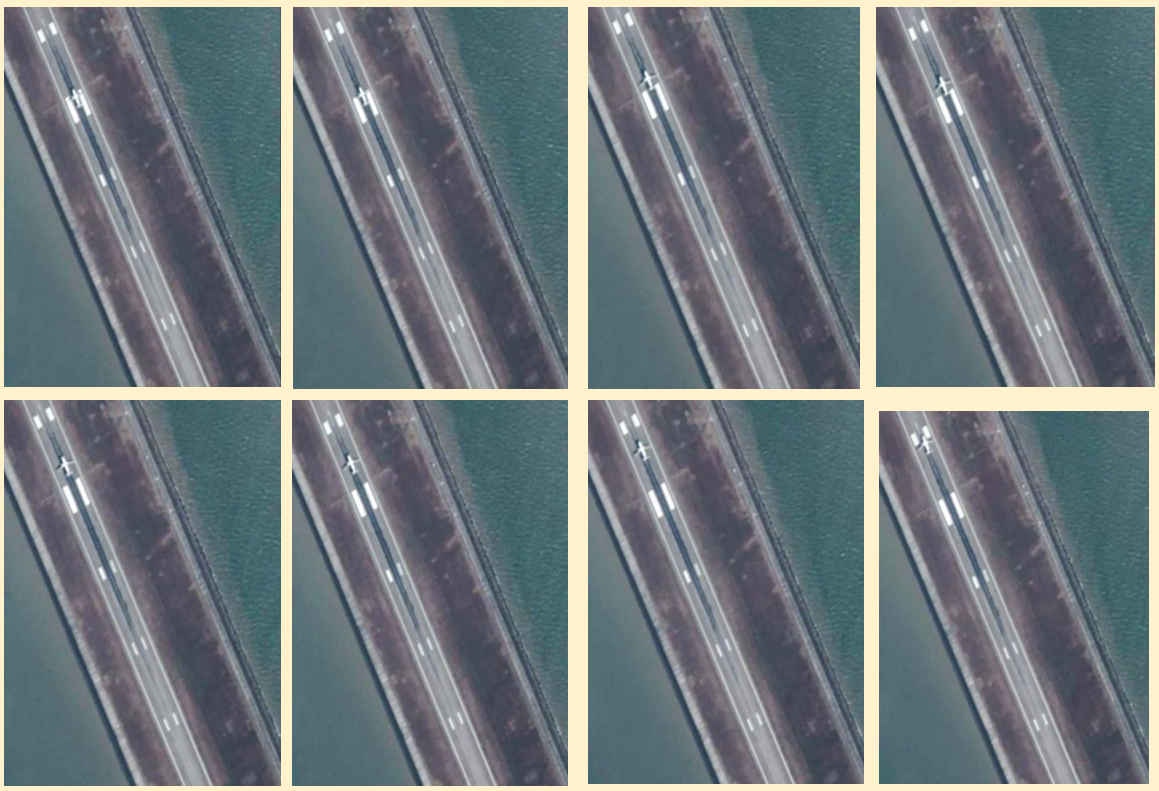
B. The aircraft flies from the upper-left region toward the lower-left region.

C. The aircraft flies from the middle of the frame toward the lower-left region.

D. The aircraft flies from the middle of the frame toward the upper-left region.

E. The aircraft flies from the lower-left region toward the lower-right region.

Figure 21. Qualitative example for Leaf 15, long-term trajectory summary for small moving targets.



Classification:

L2 Reasoning / G5 Spatiotemporal Evolution Reasoning / T12 Global Spatiotemporal Understanding / 16 Displacement and speed estimation

Question:

Based on the visible clip, what best describes the aircraft target's overall movement direction and displacement magnitude?

Options:

A. The target moves toward the upper-right, with a total displacement of about 0–20 pixels.

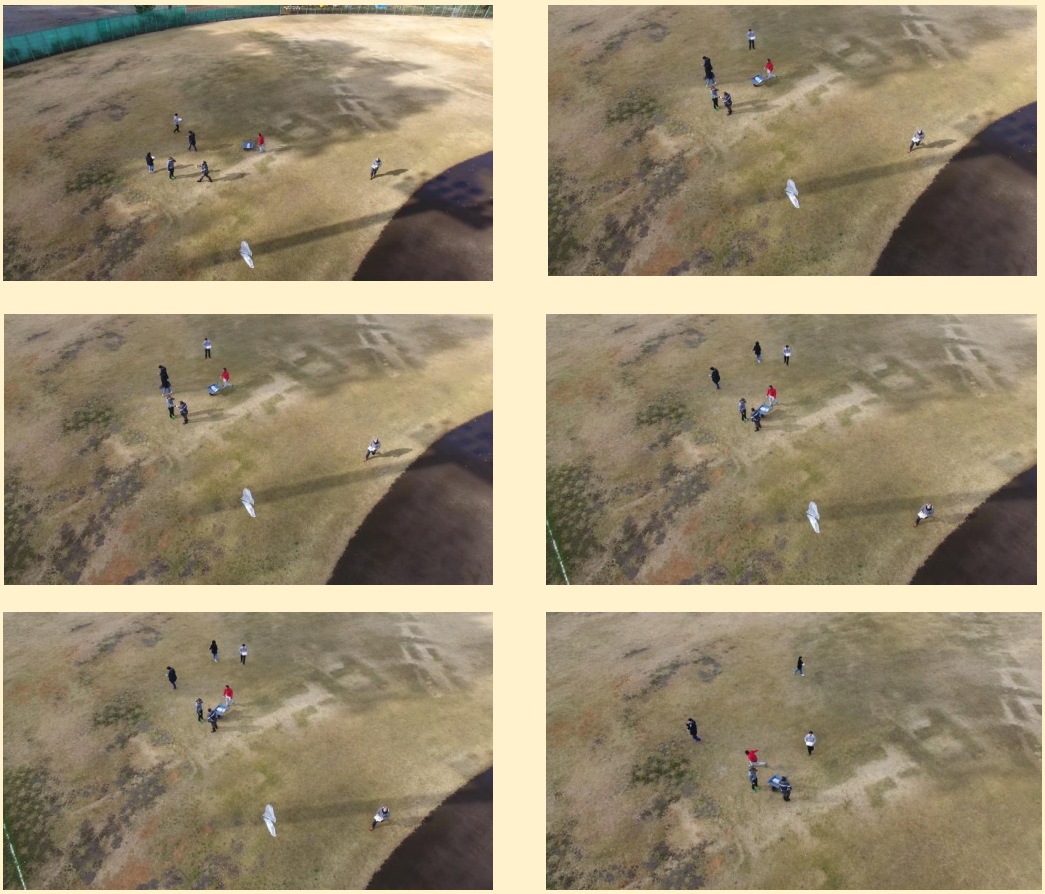
B. The target moves toward the upper-left, with a total displacement of about 60–120 pixels.

C. The target moves toward the lower-right, with a total displacement of about 20–60 pixels.

D. The target moves toward the lower-right, with a total displacement of about 120–200 pixels.

E. No moving aircraft target is seen in the visible clip.

Figure 22. Qualitative example for Leaf 16, displacement and speed estimation.



Classification:

L2 Reasoning / G5 Spatiotemporal Evolution Reasoning / T13 Local-Temporal Understanding / 17 Temporal ordering

Question:

For the red-box target, which option correctly describes the temporal order of actions across the three time intervals?

Options:

A. 0–3 s: walking; 3–5 s: running; 5–8 s: walking

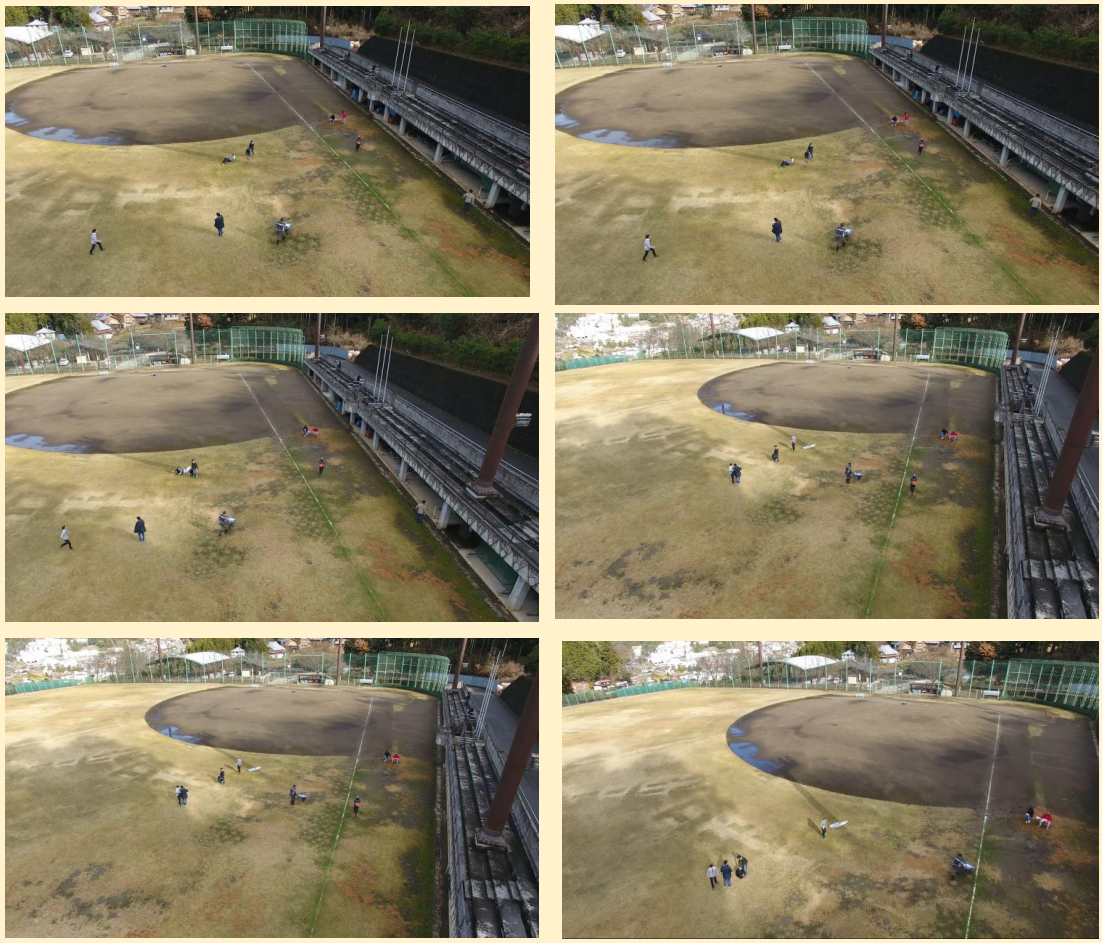
B. 5–8 s: standing; 3–5 s: shaking hands; 0–3 s: walking

C. 0–3 s: walking; 3–5 s: standing; 5–8 s: running

D. None of the choices is justified by the observed video content.

E. 0–3 s: walking; 3–5 s: shaking hands; 5–8 s: standing

Figure 23. Qualitative example for Leaf 17, temporal ordering.



Classification:

L2 Reasoning / G5 Spatiotemporal Evolution Reasoning / T13 Local-Temporal Understanding / 18 Action-duration comparison

Question:

For the red-box target, which activity occupies the longest total duration in the visible clip?

Answer:

A. Running occupies the longest total duration.

B. Hugging occupies the longest total duration.

C. Walking occupies the longest total duration.

D. Pushing or pulling occupies the longest total duration.

E. None of the listed options is supported by the visible evidence.

Figure 24. Qualitative example for Leaf 18, action-duration comparison.

Classification:

L2 Reasoning / G5 Spatiotemporal Evolution Reasoning / T13 Local-Temporal Understanding / 19 Concurrent visibility interval for multiple targets

Question:

After reviewing the clip, which option best describes the time interval during which the red and bule visually marked targets are both visible?

Options:

A. From 6 s to 12 s

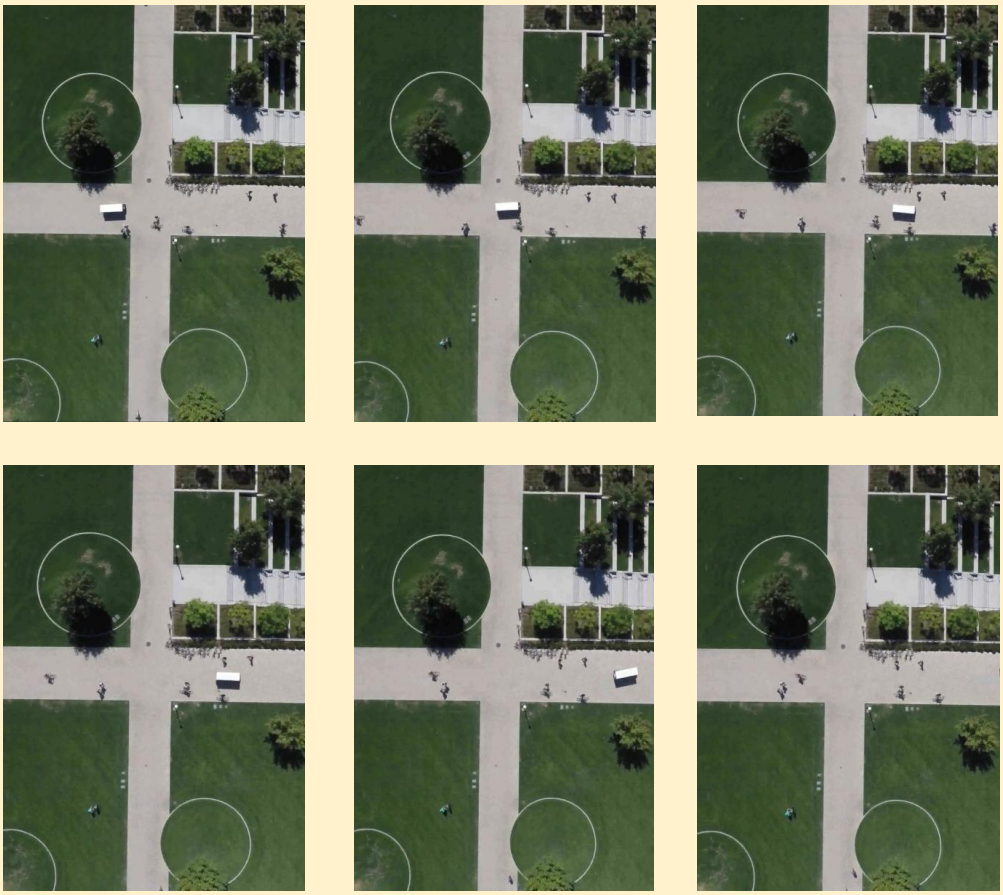
B. From 2 s to 8 s

C. From 0 s to 11 s

D. From 0 s to 6 s

E. From 10 s to 16 s

Figure 25. Qualitative example for Leaf 19, concurrent visibility interval for multiple targets.



Classification:

L2 Reasoning / G6 Spatiotemporal Localization Reasoning / T14 Single-Event Spatiotemporal Localization / 20 Single-event spatiotemporal localization

Question:

For the disappearance event of the white car visible in the video, which time interval is the closest match?

Options:

A. The supported time interval is from 2 s to 3 s.

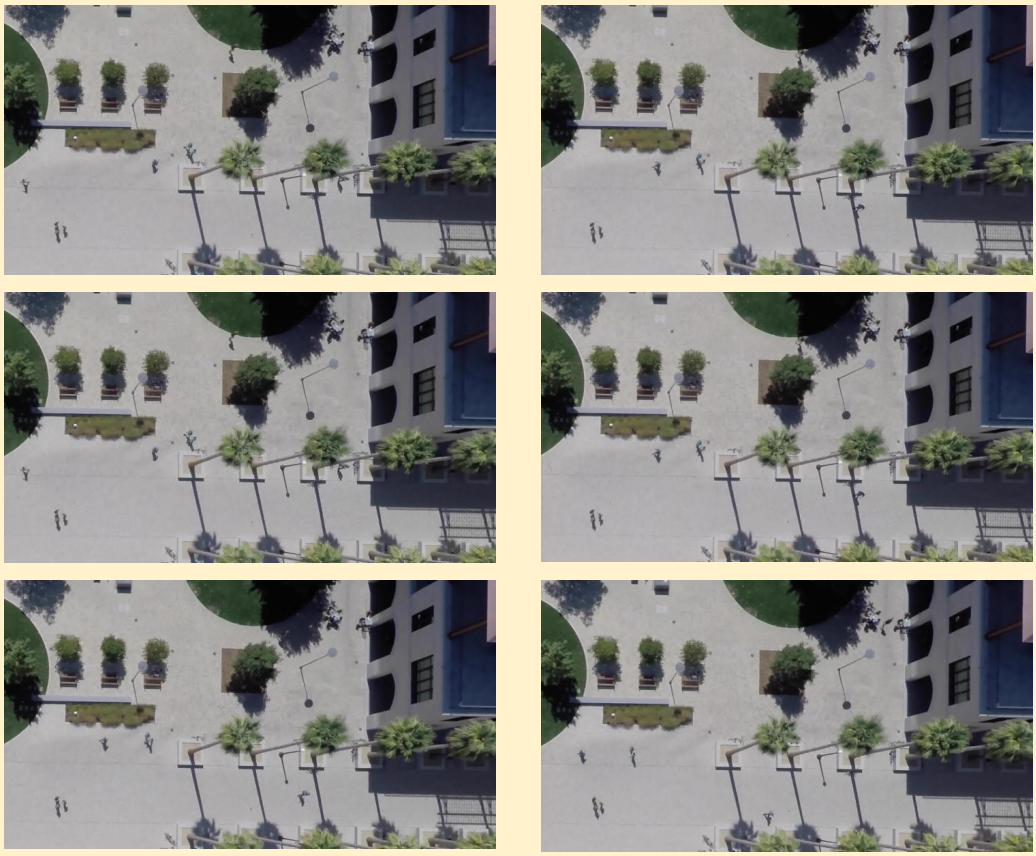
B. The supported time interval is from 3 s to 5 s.

C. The supported time interval is from 5 s to 7 s.

D. The supported time interval is from 7 s to 9 s.

E. None of the above options matches the video evidence.

Figure 26. Qualitative example for Leaf 20, single-event spatiotemporal localization.



Classification:

L2 Reasoning / G6 Spatiotemporal Localization Reasoning / T15 Multi-Event Spatiotemporal Localization / 21 Regional trajectory co-occurrence

Question:

During the specified time window, which option best describes whether the red-box and blue-box targets co-occur in the highlighted lower-center region?

Answer:

A. None of the listed options is supported by the visible evidence.

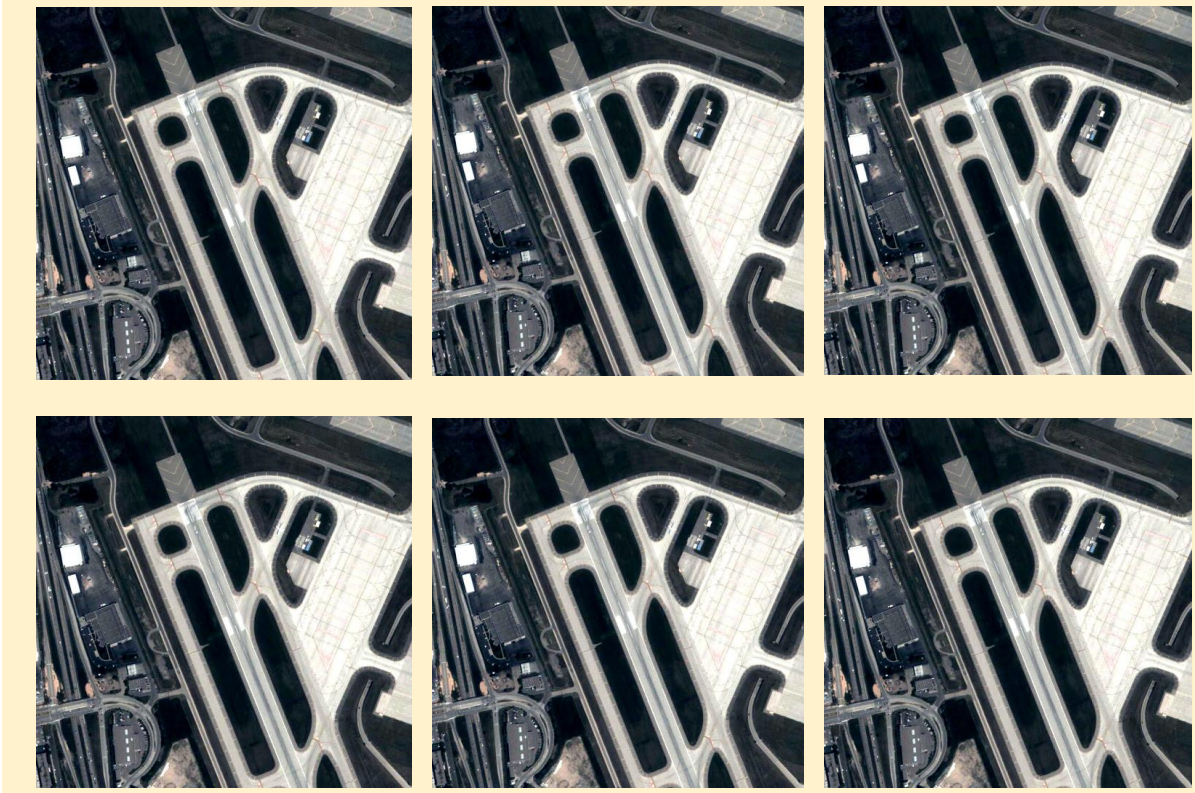
B. Only the blue-box target reaches the lower-center region during the clip.

C. Both marked targets are visible in the lower-center region at the same time around 6 s.

D. Only the blue-box target reaches the center region during the clip.

E. Both marked targets appear in the clip, but they are not visible in the lower-center region at the same time.

Figure 27. Qualitative example for Leaf 21, regional trajectory co-occurrence.



Classification:

L2 Reasoning / G7 Spatiotemporal Consistency Reasoning / T16 Spatial Relation Consistency / 22 Relation-constrained candidate localization

Question:

Based on the visible clip, in which area of the 3*3 grid is the local standby vehicle capable of carrying more than 50 people located on the runway?

Options:

A. Upper-left

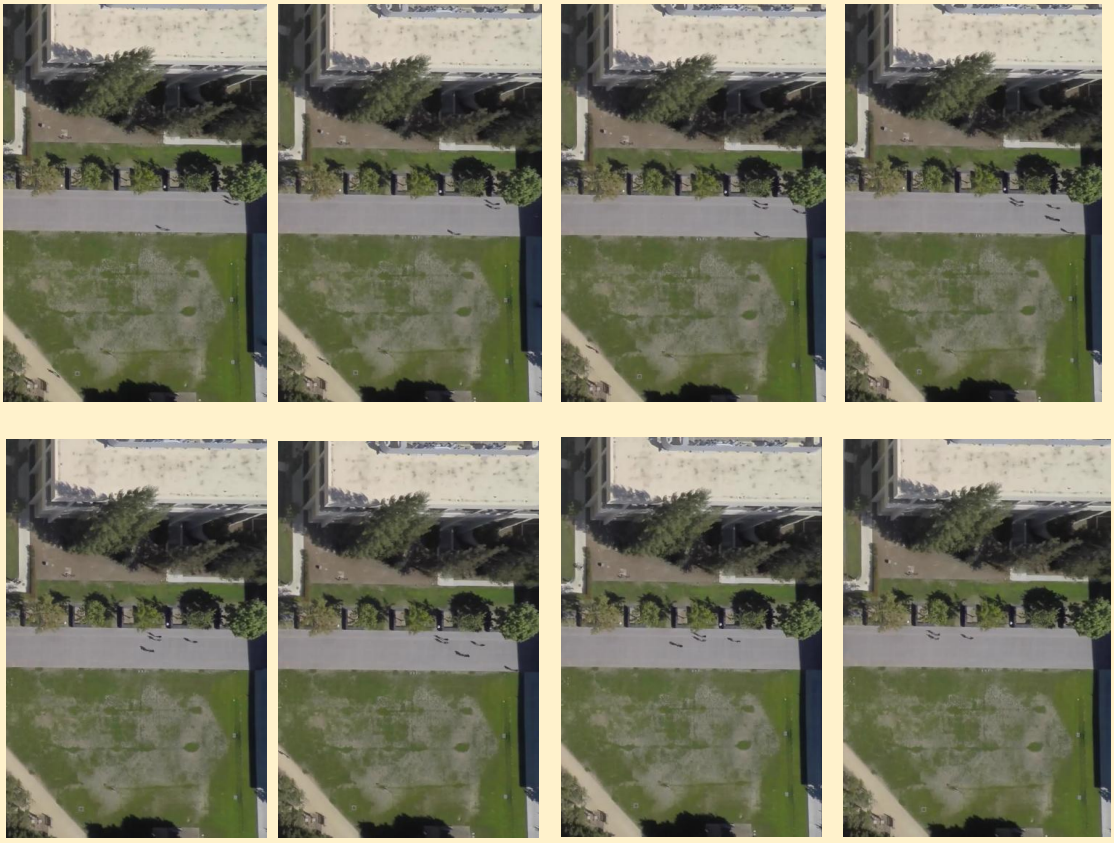
B. Upper-center

C. Middle-left

D. Center

E. None of the listed areas is supported by the visible evidence

Figure 28. Qualitative example for Leaf 22, relation-constrained candidate localization.



Classification:

L2 Reasoning / G7 Spatiotemporal Consistency Reasoning / T16 Spatial Relation Consistency / 23 Reasoned counting of small moving targets

Question:

Based on the visible evidence in the clip, how many people are located in areas shadowed by trees or buildings?

Answer:

A. 5

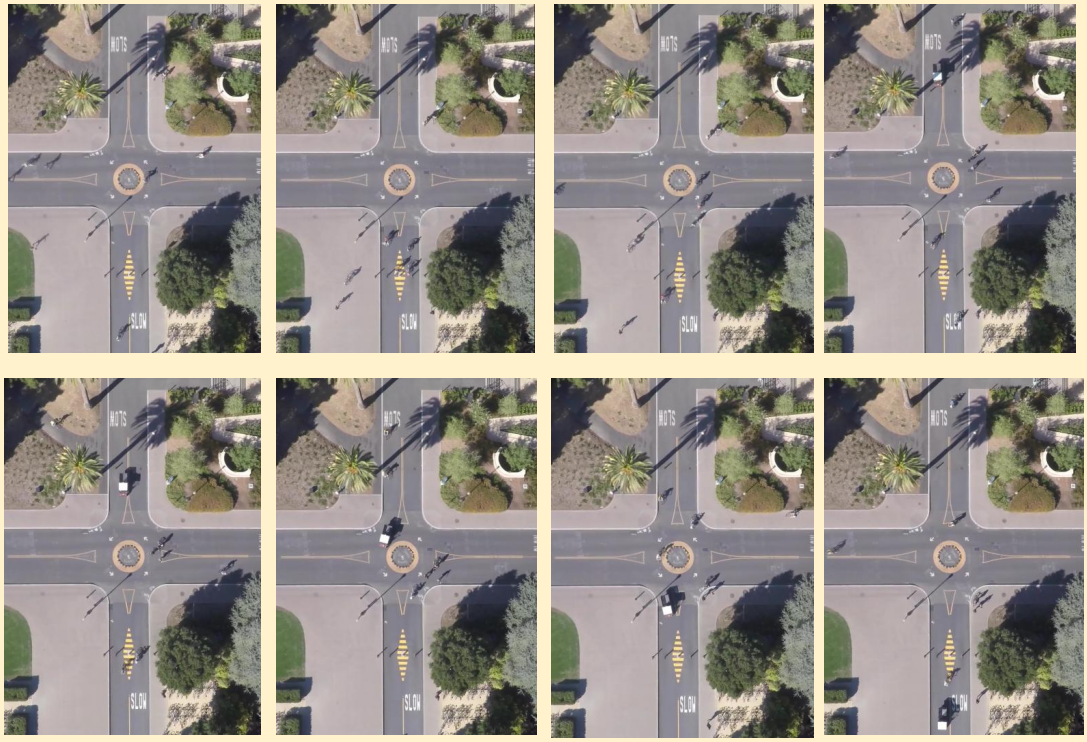
B. 6

C. 7

D. 8

E. None of the listed counts is supported by the visible evidence.

Figure 29. Qualitative example for Leaf 23, reasoned counting of small moving targets.



Classification:

L2 Reasoning / G7 Spatiotemporal Consistency Reasoning / T17 Target Consistency / 24 Cross-window constraint satisfiability

Question:

In the full video clip, is there a single cyclist who satisfies all of the following cross-window constraints: entering or exiting via the path in the upper-left part of the frame, subsequently passing through the central intersection area, and finally leaving the visible range from the lower part of the frame?

Answer:

A. The same cyclist continuously passes through all three specified regions and satisfies all constraints.

B. No cyclist enters or exits near the upper-left path.

C. A cyclist passes through the upper-left path, but does not subsequently pass through the central intersection area.

D. A cyclist passes through the first two regions, but does not leave the visible range from the lower part of the frame.

E. Cyclists appear in these regions separately, but the regions cannot be linked by a continuous trajectory of the same cyclist.

Figure 30. Qualitative example for Leaf 24, cross-window constraint satisfiability.

Classification:

L2 Reasoning / G7 Spatiotemporal Consistency Reasoning / T17 Target Consistency / 25 Target re-identification after occlusion

Question:

In temporal order, the red-box marks the earlier target and the blue-box marks the later candidate. Based on scale, position change, and motion direction, which identity-continuity judgment is most reasonable?

Options:

A. The red-box and blue-box regions should be treated as different targets because their positions are too far apart for a plausible continuous motion path.

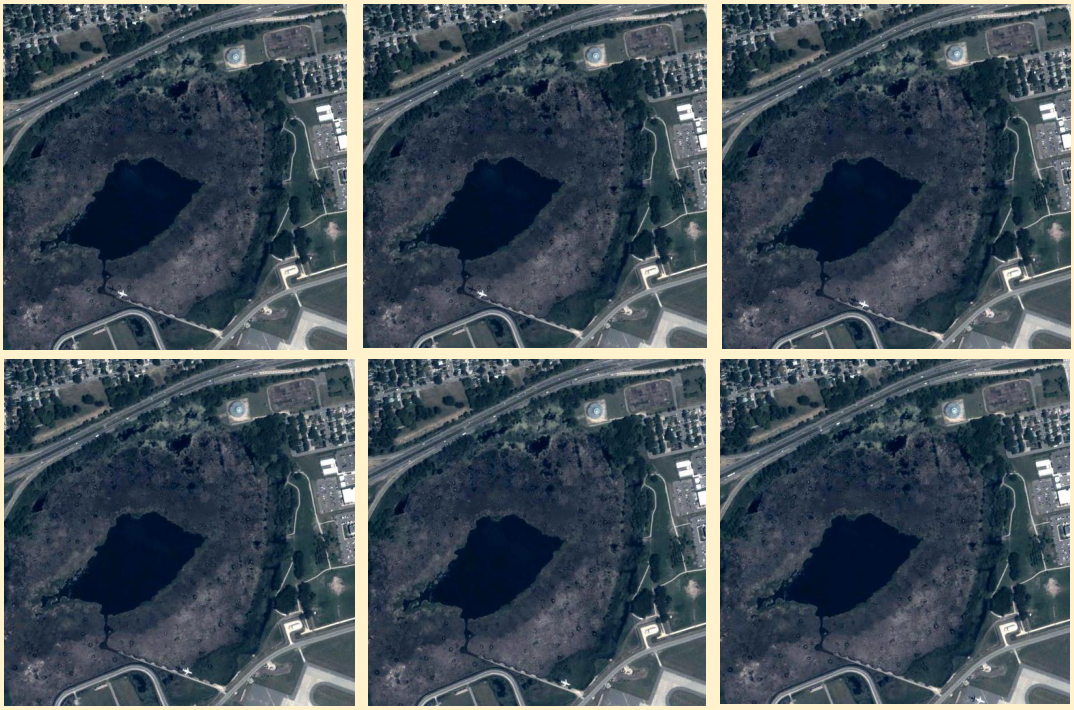
B. The blue-box marks the earlier target and the red-box marks the later candidate, so the available evidence is insufficient to link them as the same target.

C. The two marked regions should not be linked because their apparent scales and motion directions are inconsistent across the visible interval.

D. The two visually consistent, continuously visible regions are identified as the same target based on their proximity and speed-consistent motion.

E. The identity relationship cannot be determined from the marked evidence because neither region provides a stable visual cue for comparison.

Figure 31. Qualitative example for Leaf 25, target re-identification after occlusion.



Classification:

L2 Reasoning / G7 Spatiotemporal Consistency Reasoning / T17 Target Consistency / 26 Minimal constraint relaxation

Question:

The original constraint set is as follows: a single vehicle must first appear near the top side of the lake, then move toward the lower-right part of the frame, and finally leave the visible range through the lower-right corner. If no single vehicle fully satisfies all three constraints, which one constraint should be minimally relaxed so that one vehicle can satisfy the resulting constraint set?

Options:

A. Relax the first constraint from "first appearing near the top side of the lake" to "first appearing near the lakeshore or on the road adjacent to the lake."

B. Relax the motion constraint from "moving toward the lower-right part of the frame" to "moving in any visible direction."

C. Relax the final constraint from "leaving the visible range through the lower-right corner" to "leaving the visible range through any edge of the frame."

D. Remove the same-vehicle requirement and allow the three constraints to be satisfied by different vehicles.

E. No relaxation is needed, because the original constraint set is already satisfied by one vehicle.

Figure 32. Qualitative example for Leaf 26, minimal constraint relaxation.

G. Datasheets

G.1. Motivation

1. *"For what purpose was the dataset created?"*
   A: RSVideo-10K is designed to evaluate whether vision–language models can understand continuous remote-sensing videos. Existing remote-sensing benchmarks mainly focus on static images or long-interval multi-temporal observations, while general video benchmarks do not reproduce overhead viewpoints, small targets, repetitive backgrounds, and scene-constrained spatial relations. RSVideo-10K fills this evaluation gap by providing a unified five-choice video-question-answering benchmark for perception and reasoning over the included remote-sensing video sources. It supports controlled evaluation on RSVideo-Bench and method development on RSVideo-Instruct, with a taxonomy covering L1 perception and L2 reasoning over seven capability groups, 17 tasks, and 26 operational leaf capabilities.
2. *"Who created the dataset (e.g., which team, research group) and on behalf of which entity?"*
   A: The dataset is created and maintained by the following authors:
   - Hongjie Zhou, Shiqin Wang, Haoyang Chen, Haonan Guo, Di Wang, Juhua Liu, Fu Lin, and Yong Luo
3. *"Who funded the creation of the dataset?"*
   A: The funding source is not specified in the current manuscript.

G.2. Composition

Most of the questions in this section are intended to provide dataset consumers with the information they need to make informed decisions about using the dataset for their chosen tasks. Some of the questions are designed to elicit information about compliance with the EU's General Data Protection Regulation (GDPR) or comparable regulations in other jurisdictions. Questions that apply only to datasets that relate to people are grouped together at the end of the section. We recommend taking a broad interpretation of whether a dataset relates to people. For example, any dataset containing text that was written by people relates to people.

1. *"What do the instances that comprise our datasets represent (e.g., documents, photos, people, countries)?"*
   A: Each instance represents a question-specific remote-sensing video evidence state. A released record contains a video/evidence-clip identifier, one question, five answer options, the gold option, split membership, source provenance, and taxonomy labels. When needed, an instance also includes a public visual mark or temporal evidence annotation.
2. *"How many instances are there in total (of each type, if appropriate)?"*
   A: RSVideo-10K contains 10,773 five-choice question–answer instances associated with 4,629 audited evidence clips from eight public remote-sensing video sources. The dataset covers 17.02 hours of continuous video and 1,473,150 decoded frames. It includes 5,651 training items, 2,391 validation items, and 2,731 fixed test items.
3. *"Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?"*
   A: RSVideo-10K is a curated sample from eight public sources. It does not enumerate every remote-sensing platform, geographic region, weather condition, target class, scene type, or real-world event that may occur in remote-sensing videos.
4. *"Is there a label or target associated with each instance?"*
   A: Yes. Each instance has one canonical gold answer among five options, together with capability-level, group, task, and leaf labels. Records also include source and split identifiers; when applicable, they include temporal windows, spatial regions, or rendered marks that identify the observable evidence used during annotation.
5. *"Is any information missing from individual instances?"*
   A: No information required by the official five-choice evaluation interface is intentionally missing. Construction traces, adjudication notes, and train-only evidence supervision are not exposed to validation or test-time models by design.
6. *"Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)?"*
   A: Yes. Stable source, video, segment, split, task, and leaf identifiers make shared provenance and taxonomy relationships explicit. The dataset does not include social-network or user-level relationship data.
7. *"Are there recommended data splits (e.g., training, development/validation, testing)?"*
   A: Yes. RSVideo-Instruct contains 5,651 training items and 2,391 validation items. RSVideo-Bench contains 2,731 fixed test items and is reserved for final evaluation.
8. *"Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)?"*
   A: The release is license-aware and hybrid. It directly releases derived annotations, questions, answer options, taxonomy records, split identifiers, marks, evaluation code, and provenance metadata. Visual clips are redistributed only when permitted by the original source terms; otherwise the release provides source identifiers, temporal boundaries, checksums, and retrieval or reconstruction instructions.
9. *"Does the dataset contain data that might be considered*

confidential (e.g., data that is protected by legal privilege or by doctor–patient confidentiality, data that includes the content of individuals' non-public communications)?"

A: To the best of our knowledge, RSVideo-10K does not contain confidential records, privileged communications, medical records, financial records, or legal records. The visual evidence is derived from publicly available remote-sensing video sources.

10. *"Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?"*

    A: No such content is intentionally included. Some public aerial or satellite videos may show traffic, pedestrians, vehicles, emergency-related evidence, or disaster-related scene changes at remote-sensing scale. The dataset does not introduce identity labels, biometric attributes, or person-specific private information.

G.3. Collection Process

In addition to the goals outlined in the previous section, the questions in this section are designed to elicit information that may help researchers and practitioners create alternative datasets with similar characteristics. Again, questions that apply only to datasets that relate to people are grouped together at the end of the section.

1. *"How was the data associated with each instance acquired?"*

   A: The visual inputs come from eight existing public UAV, aerial, overhead, and satellite video resources. We construct the questions, options, gold answers, public marks, taxonomy assignments, and audit records for RSVideo-10K, while preserving source attribution through source and segment identifiers.
2. *"What mechanisms or procedures were used to collect the data (e.g., hardware apparatuses or sensors, manual human curation, software programs, software APIs)?"*

   A: RSVideo-10K is constructed through source registration and decoding, clip screening, task and leaf binding, spatiotemporal evidence construction, question and answer drafting, candidate construction, joint evidence–label correction, independent review, expert adjudication, split isolation, and release audit. Three senior domain experts perform initial annotation, independent review, and adjudication. Auxiliary large-model checks serve only to flag possible wording or consistency issues; human experts determine the final labels and retain only items whose answer can be verified from the released visual evidence.
3. *"If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?"*

   A: We use task-driven curation rather than probabilistic sampling. Candidate segments are retained when they contain sufficient visual evidence for one of the 17 tasks and 26 operational leaves, have usable temporal continuity, and support a unique answer under the fixed five-choice interface. Ambiguous, corrupted, duplicate, temporally overlapping, or private-metadata-dependent candidates are rejected.

G.4. Preprocessing, Cleaning, and Labeling

The questions in this section are intended to provide dataset consumers with the information they need to determine whether the raw data has been processed in ways that are compatible with their chosen tasks. For example, text that has been converted into a "bag-of-words" is not suitable for tasks involving word order.

1. *"Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)?"*

   A: Yes. Public videos are decoded, screened, segmented into evidence clips, and linked to stable source and segment identifiers. Questions and answer options are normalized to a fixed five-choice interface. Optional visual marks are rendered when a target or region cannot be communicated reliably through text alone. Records are checked for decodability, temporal continuity, answerability, option uniqueness, evidence sufficiency, mark consistency, split isolation, task labels, and leaf labels. Items that fail these checks are corrected when the released frames uniquely support the correction; otherwise they are removed.
2. *"Was the 'raw' data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)?"*

   A: The underlying source videos remain associated with their original public releases and source terms. RSVideo-10K distinguishes third-party visual content from derived benchmark records. Redistributable clips are packaged directly when permitted; restricted clips are represented by provenance metadata, temporal boundaries, checksums, and reconstruction instructions. The newly constructed questions, options, marks, taxonomy assignments, splits, and audit records are maintained as derived dataset annotations.
3. *"Is the software that was used to preprocess/clean/label the data available?"*

   A: The release package will contain the preprocessing, manifest, and evaluation resources required to reproduce the benchmark interface.

G.5. Uses

The questions in this section encourage dataset creators to distinguish supported and unsupported uses. Explicit use

boundaries help dataset consumers make informed decisions and avoid potential risks or harms.

1. *"Has the dataset been used for any tasks already?"*
   A: Yes. In this paper, RSVideo-10K is used to benchmark remote-sensing video understanding in vision–language models, analyze capability-specific failures, and train, validate, and evaluate RSVideo, our evidence-aware spatiotemporal focusing method.
2. *"Is there a repository that links to any or all papers or systems that use the dataset?"*
   A: The release repository will serve as the index for benchmark resources and reported papers or systems that use RSVideo-10K.
3. *"What (other) tasks could the dataset be used for?"*
   A: Beyond the evaluations in this paper, RSVideo-10K can support academic benchmarking of remote-sensing video perception and reasoning, controlled error analysis, video-input adaptation, spatiotemporal evidence localization, small-target temporal tracking, action and event reasoning, and method development using the training and validation splits.
4. *"Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses?"*
   A: Yes. RSVideo-10K is English-only, uses a fixed five-choice format, inherits coverage limitations from eight public sources, and has naturally imbalanced task, leaf, source, platform, scene, and target distributions. It serves as a controlled benchmark rather than a complete census of remote-sensing video phenomena.
5. *"Which tasks fall outside the intended use of the dataset?"*
   A: The intended use excludes privacy-invasive monitoring, biometric identification, operational surveillance, targeting, autonomous control, disaster-response certification, and other safety-critical decision-making without appropriate authorization, independent validation, and human oversight.

G.6. Distribution

The following answers document the distribution procedure used for internal or third-party access.

1. *"Is the dataset distributed to third parties outside of the creating entity (e.g., company, institution, organization)?"*
   A: The dataset is not yet publicly distributed. Public research access will be provided through the official release channel and will remain subject to the licenses and terms of the underlying source datasets.
2. *"How is the dataset distributed (e.g., tarball on website, API, GitHub)?"*
   A: The release will use a license-aware hybrid distribution. It will include derived annotations, data splits, public marks, evaluation code, provenance records, legally redistributable clips, and retrieval or reconstruction metadata for visual content that cannot be repackaged.
3. *"When is the dataset distributed?"*
   A: The distribution point will be the paper's official publication date.
4. *"Is the dataset distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)?"*
   A: Yes. The release manifest specifies the license for the annotations, taxonomy records, marks, split manifests, and evaluation resources created for RSVideo-10K. Third-party visual content remains subject to the licenses and terms of its original sources.
5. *"Have any third parties imposed IP-based or other restrictions on the data associated with the instances?"*
   A: Yes. The eight public source datasets retain their respective licenses, access conditions, attribution requirements, and redistribution restrictions. RSVideo-10K preserves source attribution and uses metadata/reconstruction release modes when visual redistribution is not verified.
6. *"Do any export controls or other regulatory restrictions apply to the dataset or to individual instances?"*
   A: We are not aware of export-control restrictions specific to the derived benchmark annotations. Users are responsible for complying with applicable laws, regulations, source licenses, and institutional policies in their jurisdictions.

G.7. Maintenance

The following answers document dataset maintenance and communicate the maintenance procedure to dataset consumers.

1. *"Who supports, hosts, and maintains the dataset?"*
   A: The authors of this work are the designated dataset maintainers and release hosts.
2. *"How can the owner/curator/manager of the dataset be contacted (e.g., email address)?"*
   A: The dataset maintainers can be contacted at `d_wang@whu.edu.cn` and `luoyong@whu.edu.cn`.
3. *"Is there an erratum?"*
   A: There is no erratum for the initial submission. The official release and later benchmark versions document known issues and approved corrections, if any.
4. *"How is the dataset updated (e.g., to correct labeling errors, add new instances, delete instances)?"*
   A: Dataset updates correct labeling errors, revise invalid items, improve release records, or extend the benchmark. Each update carries a version identifier and change record.

5. *"How are older versions of the dataset supported, hosted, and maintained?"*
   A: Retained version identifiers and change records trace reported results to the benchmark version used; the release manifest records the hosting policy for older packages.
6. *"If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so?"*
   A: The official release channel provides a mechanism for reporting errors and proposing corrections or extensions.