SPATIALSHERLOCK: DIAGNOSING SPATIAL INCONSISTENCIES IN AI-GENERATED VIDEOS

Anonymous authors
Paper under double-blind review

ABSTRACT

Diagnosing spatial inconsistencies in AI-generated videos requires distinguishing changes explained by camera motion, object motion, and occlusion from conflicts with the depicted scene evolution. We introduce a data pipeline that connects spatial skills reconstructed from public datasets with structured human diagnoses of generated videos. The annotations specify affected regions, required spatial corrections, and explanations, making the spatial basis of each judgment separately assessable. These references support both diagnostic reasoning supervision and component-level reinforcement-learning feedback. The pipeline curates 92,000 examples across four diagnostic tasks from more than 20,000 generated videos. Using this supervision, we train SpatialSherlock, based on Qwen3.5-9B, through supervised fine-tuning and joint reinforcement learning. The model forms local diagnoses and checks them against available intervening frames before aggregating video-level judgments and temporal evidence. SpatialSherlock achieves 81.3% accuracy on frame-pair diagnosis and 78.7% on source-video diagnosis, exceeding the strongest tested commercial baselines by 40.3 and 16.7 percentage points, respectively. It also improves on its backbone across all nine external spatial benchmarks evaluated zero-shot, with gains of 1.6–16.6 percentage points.

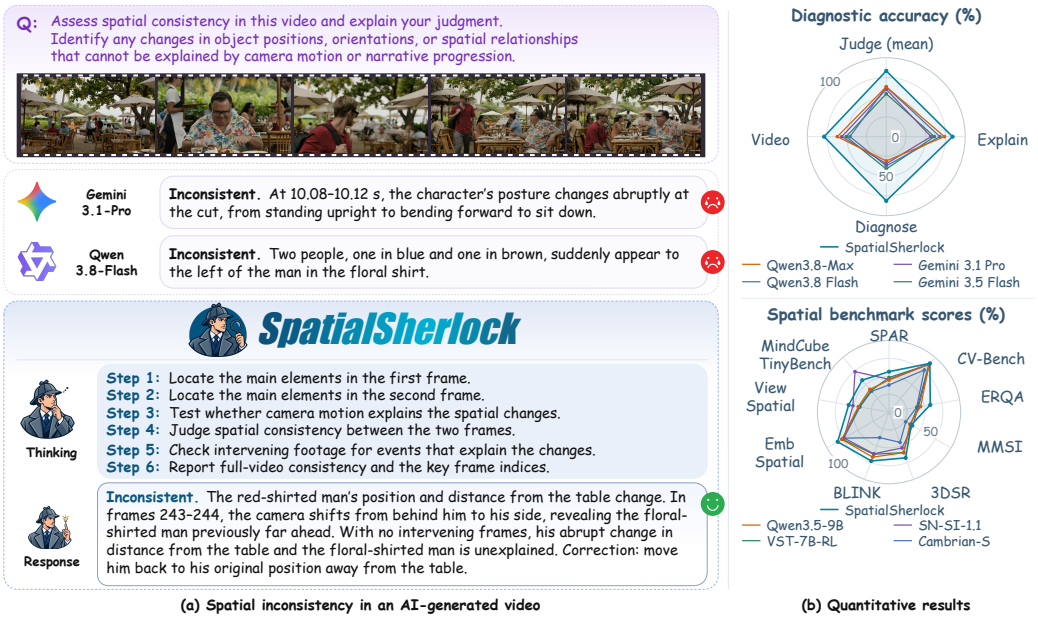

Figure 1: **SpatialSherlock diagnoses spatial inconsistencies from scene evolution.** (a) An AI-generated video with model responses; SpatialSherlock's full-video reasoning and final diagnosis appear in Appendix A. Displayed answers are condensed from the output JSON `response` field (Appendix B). Red crying and green smiling faces indicate incorrect and correct diagnoses, respectively. (b) Diagnostic accuracy and spatial benchmark scores (Tables 1 and 2).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

1 INTRODUCTION

Advances in multimodal models have improved video question answering, temporal reasoning, and the use of spatial information (Feng et al., 2025; Yang et al., 2026b). These capabilities bring video understanding closer to a broader goal: understanding the depicted world and how it evolves. AI-generated videos pose a demanding test because convincing imagery and smooth motion can coexist with spatial inconsistencies (Zheng et al., 2025; Dou et al., 2026). Diagnosing these inconsistencies identifies what needs correction, providing a basis for evaluating and improving generated videos. The central difficulty is attribution: camera motion, object motion, and occlusion can all change apparent spatial relationships and visibility. A useful diagnosis must establish which observed changes conflict with the scene evolution supported by video evidence.

Recent models organize spatial observations through geometrically supervised textual reasoning traces (Yang et al., 2026a), cognitive maps (Ouyang et al., 2025), and depth- and pose-aware memory with cross-view consistency training (Huang et al., 2026). These representations support spatial reasoning; diagnosis additionally requires assessing whether observed events justify changes in spatial relationships. In physical-violation diagnosis, PhyDetEx jointly learns judgments and free-form explanations, including for object penetration, and evaluates each explanation as a whole (Wang et al., 2025b). Physion-Eval provides human rationales and temporal annotations, and its qualitative analyses of video critics report misplaced temporal evidence and explanations unsupported by observed events (Zhang et al., 2026). A judgment or an overall explanation score offers limited guidance on which visual claims underlying a diagnosis need correction. Targeted feedback therefore calls for diagnostic annotations that make these supporting claims separately assessable against visual evidence.

Our approach connects the spatial skills needed for attribution with supervision of the diagnosis itself. We reconstruct object localization, relative-position reasoning, and camera-motion examples from public datasets, then combine them with human diagnostic annotations of generated videos. The skill examples provide a perceptual foundation; the diagnostic annotations connect visual evidence to consistency judgments. These annotations identify affected regions, specify required spatial corrections, and explain whether observed changes are compatible with the motion and visibility supported by the video. Corrections make the expected spatial state explicit by stating what should be visible and where. Each component has a human reference and can be assessed separately, so the same annotations support reasoning supervision and graded reinforcement-learning feedback. From more than 20,000 generated videos, the pipeline curates 92,000 task examples spanning Judge, Explain, Diagnose, and Video. Using this supervision, we train SpatialSherlock from Qwen3.5-9B through supervised fine-tuning and joint reinforcement learning (Figure 1). At inference, the model forms local diagnoses and, when intervening frames are available, uses the same backbone to verify or revise them against observed events. It aggregates the resulting diagnoses into a source-video judgment, explanation, and supporting temporal evidence.

SpatialSherlock outperforms the tested commercial models across all four diagnostic tasks (Table 1). On Diagnose, which jointly predicts a judgment and explanation from frame pairs, accuracy reaches 81.3%, compared with 41.0% for Gemini 3.5 Flash. Video evaluates source-video judgments, explanations, and temporal evidence; accuracy reaches 78.7%, compared with 62.0% for Qwen3.8-Max. The model also improves on its Qwen3.5-9B backbone across all nine external spatial benchmarks evaluated zero-shot, with gains of 1.6–16.6 percentage points. It ranks highest among the compared models on seven of these benchmarks (Table 2).

Our contributions are threefold. **(1) A data pipeline for spatial diagnosis.** We connect reconstructed spatial skills with structured human diagnoses, providing separately assessable references for reasoning supervision and component-level feedback. **(2) A model that diagnoses and verifies spatial inconsistencies.** SpatialSherlock learns from this supervision and checks local diagnoses against available intervening events before producing a video-level conclusion. **(3) Evaluation of diagnosis and broader spatial understanding.** Four diagnostic tasks assess judgments, explanations, their joint prediction, and temporal evidence; nine external benchmarks assess broader spatial capabilities.

2 RELATED WORK

2.1 AI-GENERATED VIDEOS

Commercial video generators such as Seedance (BytePlus, 2026) and HappyHorse (Mo, 2026) are already used in short-form drama production, brand advertising, and e-commerce marketing. Evaluation spans subject consistency and motion smoothness in VBench (Huang et al., 2024), physics and commonsense in VBench-2.0, and camera-pose agreement in Spatial Geometric Consistency. For production quality checks, spatial diagnosis identifies inconsistent relationships and explains what needs correction.

2.2 VIDEO AND SPATIAL UNDERSTANDING MODELS

Spatial understanding models acquire scene relationships through geometric supervision and structured representations. SpatialVLM scales spatial question answering (Chen et al., 2024); SpatialRGPT combines 3D scene-graph supervision with depth cues (Cheng et al., 2024). Visual Spatial Tuning (VST) learns geometrically supervised reasoning, SpaceR uses cognitive maps, and ConsiSpace combines depth- and pose-aware memory with cross-view consistency training. We connect reconstructed spatial skills with human diagnostic annotations so that models learn both consistency judgments and the spatial grounds for them.

Diagnostic and evidence-grounded models further connect judgments to visual evidence. PhyDetEx jointly learns physical-plausibility judgments and explanations, while Physion-Eval supplies human rationales and temporal annotations. DeeptraceReward supervises regions, onset times, and explanations of generation traces (Fu et al., 2025). Open-o3-Video rewards answer correctness, temporal alignment, and spatial grounding (Meng et al., 2025). Our annotations additionally specify required spatial corrections, making explicit what should be visible and where. Regions, corrections, and incompatibility explanations provide separately assessable references for reasoning supervision and graded feedback.

Video-R1 reinforces temporal reasoning, while Cambrian-S develops prediction-driven memory and event segmentation. Open-o3-Video also scores cropped visual evidence to weight candidate answers at inference. SpatialSherlock targets a specific verification question: can intervening events explain a locally diagnosed spatial discrepancy? Conditioned on the local diagnosis, it uses available intervening frames to verify or revise the judgment against observed camera motion, object motion, and occlusion.

3 DATA PIPELINE FOR SPATIAL DIAGNOSIS

We define a spatial inconsistency as a conflict between observed spatial relationships and the scene evolution supported by video evidence. Our data pipeline (Figure 2) combines reconstructed spatial-skill examples with human diagnostic annotations of generated videos. These annotations connect judgments to spatial evidence for reasoning supervision and component-level feedback.

3.1 RECONSTRUCTING SUPERVISION FOR SPATIAL SKILLS

We reconstruct spatial-skill examples primarily from HSSD (Khanna et al., 2024), MS-COCO (Lin et al., 2014), ScanNet (Dai et al., 2017), ScanNet++ (Yeshwanth et al., 2023), and ARKitScenes (Baruch et al., 2021). We use Qwen3.8-Max (Qwen Team, 2026) to construct questions, answers, and reasoning from official images and annotations, including camera poses and 3D geometry where available. The targets cover object localization, relative-position reasoning within and across images, and camera motion. These skills support the spatial observations and comparisons needed to explain apparent changes. The resulting mixed pool contains approximately 200,000 spatial-skill examples. Sources and construction are detailed in Appendix C.

3.2 CONSTRUCTING GENERATED-VIDEO DIAGNOSTIC SUPERVISION

Generating videos and constructing intervals. Using shot-by-shot scripts from short-drama and advertising production, we generate over 20,000 videos of 10–30 s with multiple video generators,

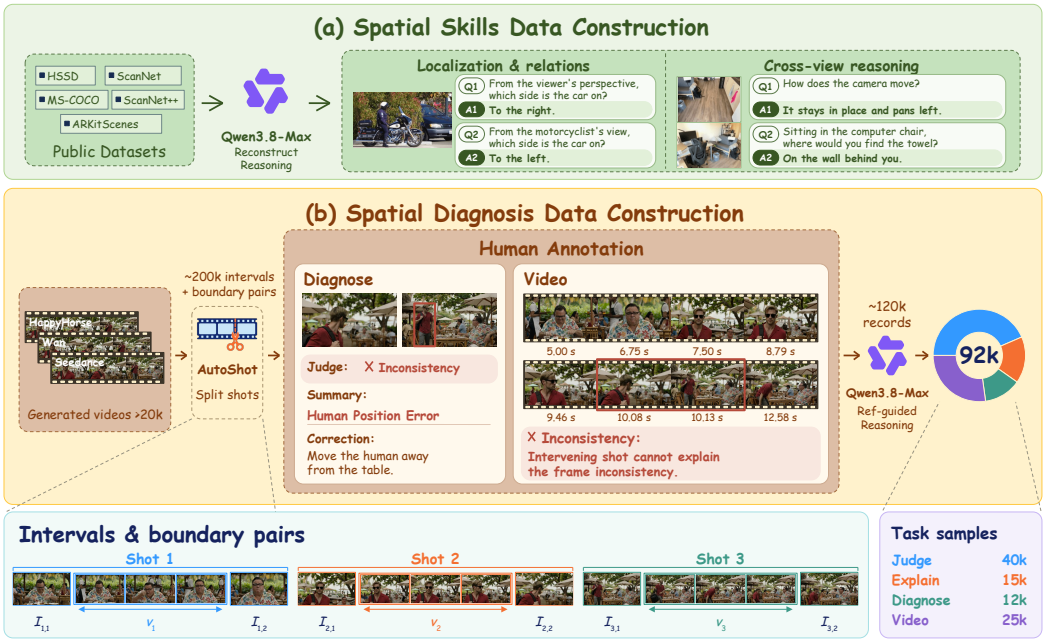

Figure 2: **Data pipeline for spatial diagnosis.** (a) Spatial-skill supervision from public annotations. (b) Generated-video supervision from human diagnoses and reference-guided reasoning. Lower insets show shot segmentation and task sample counts.

such as Seedance and HappyHorse. AutoShot (Zhu et al., 2023) detects visual discontinuities and partitions each source video V into n basic segments:

$$S_i = \big[I_{i,1};\, v_i;\, I_{i,2}\big], \qquad V = S_1 \,\|\, \cdots \,\|\, S_n,$$

where $I_{i,1}$ and $I_{i,2}$ are the first and last frames of segment i, v_i contains the visually continuous intervening frames, and $\|$ denotes temporal concatenation.

Because intervening events may explain endpoint discrepancies, we retain every contiguous combination of segments as a candidate interval:

$$C_{ij} = S_i \,\|\, \cdots \,\|\, S_j, \qquad 1 \le i \le j \le n.$$

The $n(n+1)/2$ intervals preserve all intervening frames and transitions, providing event context at multiple temporal scales. We separately retain the $n-1$ boundary pairs $B_i = (I_{i,2};\, I_{i+1,1})$, which isolate changes across adjacent segments. This yields approximately 200,000 candidate intervals, plus the boundary pairs.

Annotating and verifying diagnoses. Qwen3.8-Max first filters visually degraded or uninformative pairs and guides trimming of the corresponding shots. Human annotators then diagnose each retained endpoint or boundary pair E. Its local record contains y_E, discrepancy regions, required spatial corrections, and an explanation grounded in observed motion and visibility. Corrections specify the expected spatial state, making the judgment's spatial assumptions explicit. For each retained C_{ij}, annotators then verify the endpoint diagnosis using camera motion, object motion, and occlusion in the intervening frames. The interval-level record contains y_C, an explanation, and supporting timestamps or intervals; this wider evidence can change the judgment. Both labels are binary: 1 denotes a supported inconsistency within the respective evidence scope and 0 its absence. Boundary pairs contribute only local records; the combined pool contains approximately 120,000 records.

Curating task supervision. Reference-guided reasoning construction and a second screening yield 92,000 task examples (Figure 2, lower right). Judge predicts y_E; Explain generates an explanation conditioned on the supplied y_E. Diagnose jointly predicts the judgment and explanation, with

region–correction references supporting component-level assessment. Video training uses interval judgments, explanations, and temporal evidence; evaluation assesses the aggregated source-video diagnosis. All examples derived from a source video remain in the same data partition.

Guiding reasoning with diagnostic references. We prompt Qwen3.8-Max with each annotated example's visual input, reference judgment, and human diagnostic annotations to construct step-by-step reasoning. The prompt requires the reasoning to connect visual observations to the judgment using annotated region–correction pairs, explanations, and task-specific temporal evidence. During generation, the teacher flags suspected annotation errors for human review. Appendix C details the two screening stages, reasoning construction, and data partitioning.

4 SpatialSherlock: Training and Inference

SpatialSherlock trains a Qwen3.5-9B backbone through supervised fine-tuning (SFT) and joint reinforcement learning (RL). The RL stage uses group relative policy optimization (GRPO) (Shao et al., 2024) with the DAPO recipe (Yu et al., 2025) in verl. Diagnostic references supply both reasoning targets and component-level rewards (Figure 3).

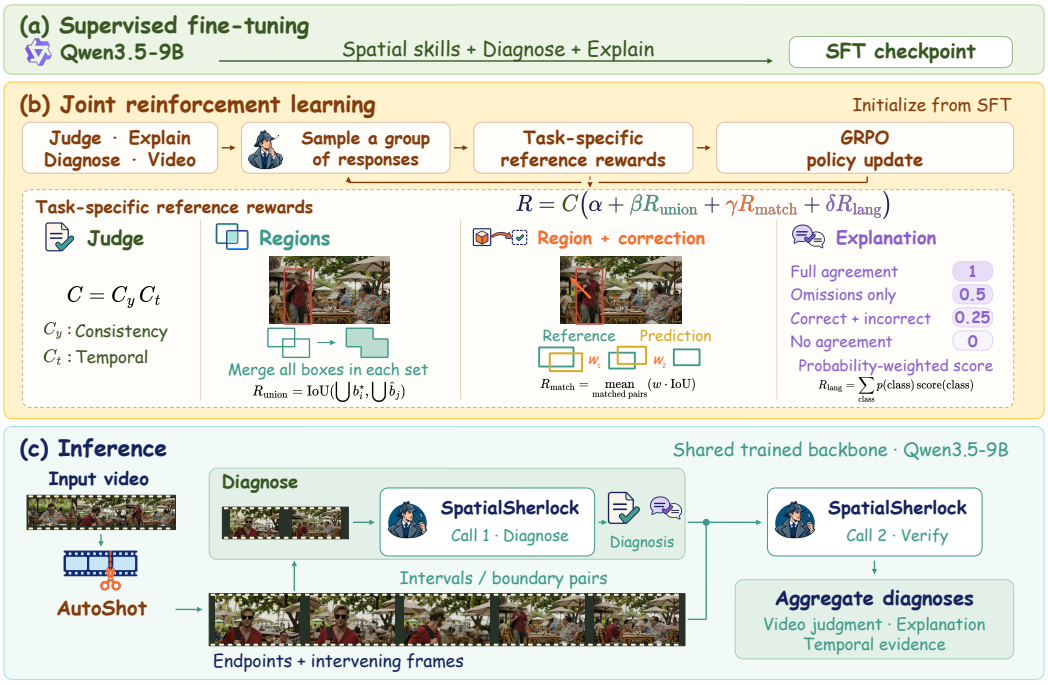

Figure 3: **SpatialSherlock training and inference.** (a) SFT on spatial skills, Diagnose, and Explain. (b) Joint RL with component-level rewards, illustrated for nonempty reference regions. (c) Local diagnosis, event verification, and video-level aggregation. Boxes and arrows are illustrative.

4.1 Training

Supervised fine-tuning. SFT connects spatial perception to diagnosis by mixing spatial-skill examples (Section 3.1) with Diagnose and Explain (Section 3.2). For visual input and instruction x and reference response $a^\star = (a_1^\star, \ldots, a_L^\star)$, we supervise reasoning and task outputs together by minimizing

$$\mathcal{L}_{\text{SFT}}(\theta) = -\mathbb{E}_{(x,a^\star)\sim\mathcal{D}_{\text{SFT}}}\left[\sum_{t=1}^{L} \log \pi_\theta(a_t^\star \mid x, a_{<t}^\star)\right]. \tag{1}$$

Joint reinforcement learning. From the SFT checkpoint, we apply GRPO jointly to Judge, Explain, Diagnose, and Video. Reference-based rewards give group-relative advantages for responses to each input. We maximize the clipped objective to favor higher-reward responses:

$$\mathcal{J}_{\mathrm{GRPO}}(\theta) = \mathbb{E}_{\mathcal{Q}}\left[\frac{1}{N_{\mathcal{Q}}}\sum_{(x,g,t)\in\mathcal{Q}} \min\{\rho_{g,t}A_g, \mathrm{clip}(\rho_{g,t}, 1-\epsilon_{\mathrm{l}}, 1+\epsilon_{\mathrm{u}})A_g\}\right]. \tag{2}$$

Here $\mathcal{Q}$ contains valid response tokens, $N_{\mathcal{Q}} = |\mathcal{Q}|$, and A_g is response g's normalized group-relative reward. Clipping the current-to-rollout probability ratio $\rho_{g,t}$ with parameters $\epsilon_{\mathrm{l}}, \epsilon_{\mathrm{u}}$ limits incentives for large updates (Appendix D).

4.2 Component-level rewards

Correctness gates the component scores so that supporting details cannot compensate for a wrong judgment. For joint diagnosis with nonempty reference regions,

$$R = C\left(\alpha + \beta R_{\mathrm{union}} + \gamma R_{\mathrm{match}} + \delta R_{\mathrm{lang}}\right), \qquad \alpha+\beta+\gamma+\delta = 1, \tag{3}$$

where nonnegative weights combine base correctness, region coverage, matched corrections, and explanation agreement. The output fields and their reward mapping are specified in Appendix B: `answer` and `video_answer` provide correctness, `bboxes` provides regions and corrections, and `summary` provides the concise diagnostic explanation. Appendix E details task-specific rewards, empty-region cases, and scoring rules.

Consistency and temporal correctness. A correct video label can be paired with an unrelated moment. To bind judgments to their supporting events, the correctness gate requires correct labels and temporal selections:

$$C = C_y C_t, \qquad C_y = \mathbf{1}[\hat{y} = y], \qquad C_t = \mathbf{1}[\hat{z} = z]. \tag{4}$$

Here $y, \hat{y}$ are reference and predicted labels; $z, \hat{z}$ index the reference and selected candidate frame-pair combinations for Video. Non-video inputs have one temporal candidate, giving $C_t = 1$. Video presents temporal candidates as a multiple-choice question: selecting the reference option earns one point and any other answer earns zero. For Video, multiplying the two scores makes both decisions necessary for a positive reward. The base term αC then credits the correct decision itself.

Regions. Intersection over union (IoU) between the reference and predicted region unions penalizes missed and excess area while counting overlapping regions once. For n reference boxes $b_i^\star$ and m predicted boxes $\hat{b}_j$,

$$U^\star = \bigcup_{i=1}^{n} b_i^\star, \qquad \hat{U} = \bigcup_{j=1}^{m} \hat{b}_j, \qquad R_{\mathrm{union}} = \frac{|U^\star \cap \hat{U}|}{|U^\star \cup \hat{U}|}. \tag{5}$$

Identical unions receive the same score even when their box decompositions differ, keeping area coverage separate from the pairing evaluated below.

Region + correction. An accurately localized object can still receive an incorrect proposed correction. Correction agreement therefore weights matched IoU, tying each action to its region. One-to-one Hungarian matching $M^\star$ maximizes total IoU; the mean includes zero-scoring pairs so that failed corrections lower the score:

$$R_{\mathrm{match}} = \frac{1}{|M^\star|}\sum_{(i,j)\in M^\star} \omega_{ij}\,\mathrm{IoU}(b_i^\star, \hat{b}_j). \tag{6}$$

For reference and predicted keyword sets $K_i^\star$ and $\hat{K}_j$, the weight gives partial credit for correct subsets and zero for any unsupported action:

$$\omega_{ij} = \begin{cases} \dfrac{|\hat{K}_j \setminus \{\texttt{move}\}| + \mathbf{1}[\texttt{move} \in \hat{K}_j] d_{ij}}{|K_i^\star|}, & K_i^\star \neq \emptyset,\ \hat{K}_j \subseteq K_i^\star, \\ 0, & \text{otherwise.} \end{cases} \tag{7}$$

Keywords are `add`, `delete`, `rotate`, `replace`, `move`, and `background`. Only `move` receives direction weight $d_{ij} = \max(0, \cos\theta_{ij})$ for nonzero, equal-dimensional vectors: angular deviations reduce credit, and orthogonal or opposing directions receive none. Geometric matching fixes the region correspondence before correction semantics are scored. Thus, moving the correct object in the wrong direction loses correction credit even when its predicted box overlaps the reference.

Explanation. The explanation reward assesses whether the diagnosis states the spatial facts and evidence given in the human reference. Qwen3.5-9B compares the predicted and reference `summary` fields, allowing semantically equivalent wording to receive credit. It assigns normalized probabilities p_k to four classes: full agreement, omissions only, mixed correct and incorrect claims, and no complete matching fact. The expected score retains all four class probabilities for graded feedback, penalizing errors more than omissions:

$$R_{\text{lang}} = \sum_{k=1}^{4} \lambda_k p_k, \qquad (\lambda_1, \lambda_2, \lambda_3, \lambda_4) = (1, 0.5, 0.25, 0). \tag{8}$$

The lower weight for mixed claims reduces the reward for adding unsupported details to an otherwise correct explanation.

4.3 Inference

Intervening events can explain endpoint discrepancies, so inference tests local diagnoses against this evidence. AutoShot boundaries yield C_{ij} and B_i (Section 3.2). The trained backbone first predicts a judgment and spatial explanation for each endpoint or boundary pair. When intervening frames are available, a second call to the same backbone uses them with the local diagnosis to verify or revise it against observed camera motion, object motion, and occlusion. Pairs without intervening frames retain their local diagnoses. The final diagnoses are aggregated into a source-video judgment, explanation, and supporting temporal evidence. AutoShot settings, 24-frame sampling, and the verification prompt are described in Appendix D.

5 Experiments

We assess diagnostic accuracy and generalization to external spatial benchmarks. Evaluation settings are in Appendices F and G.

5.1 Four-task spatial diagnosis

Judge, Explain, and Video use randomly sampled test examples; Diagnose uses a random sample of cases flagged as challenging during human annotation. SpatialSherlock led all four tasks (Table 1). Source-video accuracy reached 78.7%, exceeding the strongest commercial baseline by 16.7 percentage points. On frame pairs, it exceeded the strongest baselines by 40.3 points on Diagnose and 10.3 points on Explain.

Judge revealed opposing class preferences: both Gemini models favored negatives (inconsistent), whereas both Qwen3.8 models favored positives (consistent). Their class gaps ranged from 34.5 to 76.5 points. SpatialSherlock achieved 83.3% mean accuracy, 20.0 points above the strongest baseline, with a class gap of only 3.5 points.

5.2 Generalization experiments

We evaluated SpatialSherlock and its Qwen3.5-9B backbone zero-shot on SPAR-Bench (Zhang et al., 2025), CV-Bench (Tong et al., 2024), ERQA (Gemini Robotics Team et al., 2025), MMSI-Bench (Yang et al., 2025), 3DSRBench (Ma et al., 2025), BLINK (Fu et al., 2024), EmbSpatial-Bench (Du et al., 2024), ViewSpatial-Bench (Li et al., 2025), and MindCube (Wang et al., 2025a).

SpatialSherlock improved on the backbone across all nine benchmarks by 1.6–16.6 percentage points (Table 2), extending its gains to spatial relations and reasoning across views. Among the compared models, it led on seven benchmarks; SenseNova-SI-1.1 (SN-SI-1.1) (Cai et al., 2026) led on MMSI-Bench and MindCube TinyBench.

Table 1: Diagnostic accuracy (%). Video is scored per source video. Judge: negative = inconsistent, positive = consistent; Mean averages both classes; bold marks column maxima.

Model	Video	Diagnose	Judge			Explain
			Negative	Positive	Mean	
Gemini 3.1 Pro	56.3%	34.3%	86.5%	33.5%	60.0%	56.7%
Gemini 3.5 Flash	50.7%	41.0%	**92.5%**	16.0%	54.3%	61.0%
Qwen3.8 Flash	46.0%	37.0%	39.0%	**87.0%**	63.0%	68.0%
Qwen3.8-Max	62.0%	30.7%	46.0%	80.5%	63.3%	73.7%
SpatialSherlock	**78.7%**	**81.3%**	81.5%	85.0%	**83.3%**	**84.0%**

Table 2: External spatial scores (%); bold marks row maxima. * marks benchmark-specific training: SN-SI-1.1 uses MindCube-10K for training and disjoint TinyBench for evaluation.

Benchmark	Qwen3.5-9B	SpatialSherlock	SN-SI-1.1	VST-7B-RL	Cambrian-S
SPAR-Bench	44.8	**56.5**	45.8	48.3	37.9
CV-Bench (overall)	87.3	**88.9**	88.2	86.5	76.9
ERQA	44.8	**58.5**	41.8	40.3	39.2
MMSI-Bench	33.6	35.9	**38.1**	34.8	27.1
3DSRBench	60.3	**68.0**	53.2	60.1	45.0
BLINK (validation)	66.8	**72.6**	61.9	62.6	37.9
EmbSpatial-Bench	76.1	**82.8**	72.5	75.5	72.8
ViewSpatial-Bench	43.6	**57.2**	51.2	42.4	41.3
MindCube TinyBench	41.6	58.2	**73.8***	39.1	37.9

5.3 Qualitative visualization

Figure 4 presents four video examples, including two spatially inconsistent and two spatially consistent cases.

Ablations. Our ablations assess the contributions of spatial-skill and diagnostic data and compare joint RL with continued SFT under matched training conditions. Configurations, metrics, and results are collected in Appendix H.

6 Conclusion

We develop a data pipeline that supervises spatial diagnosis through separately assessable regions, correction descriptions, and explanations. These components connect spatial changes to consistency judgments and allow structured annotations to support both reasoning supervision and graded reinforcement-learning feedback. Trained on this supervision, SpatialSherlock carries this design into inference: local diagnosis identifies the spatial claim, and temporal verification checks it against the events connecting the observations. The resulting model achieves strong frame-pair and full-video diagnostic accuracy and improves over its backbone on nine external spatial benchmarks. The contribution is a data-driven approach to video spatial reasoning that supervises both the consistency judgment and the diagnostic basis supporting it.

AI use statement

Writing and polishing: OpenAI Codex assisted drafting, translation, and language revision. **Retrieval and discovery:** OpenAI Codex helped find related work. **Synthetic dataset generation:** Video models generated diagnostic source videos; Qwen3.8-Max constructed spatial-skill examples and annotation-guided reasoning and answers. Generation, filtering, and human review are

detailed in Section 3 and Appendix C. The authors are responsible for the data, analyses, and final manuscript.

Reproducibility Statement

Data and splits follow Section 3 and Appendix C; training, inference, and rewards follow Section 4 and Appendices D and E. Evaluation and ablation protocols are in Appendices F–H.

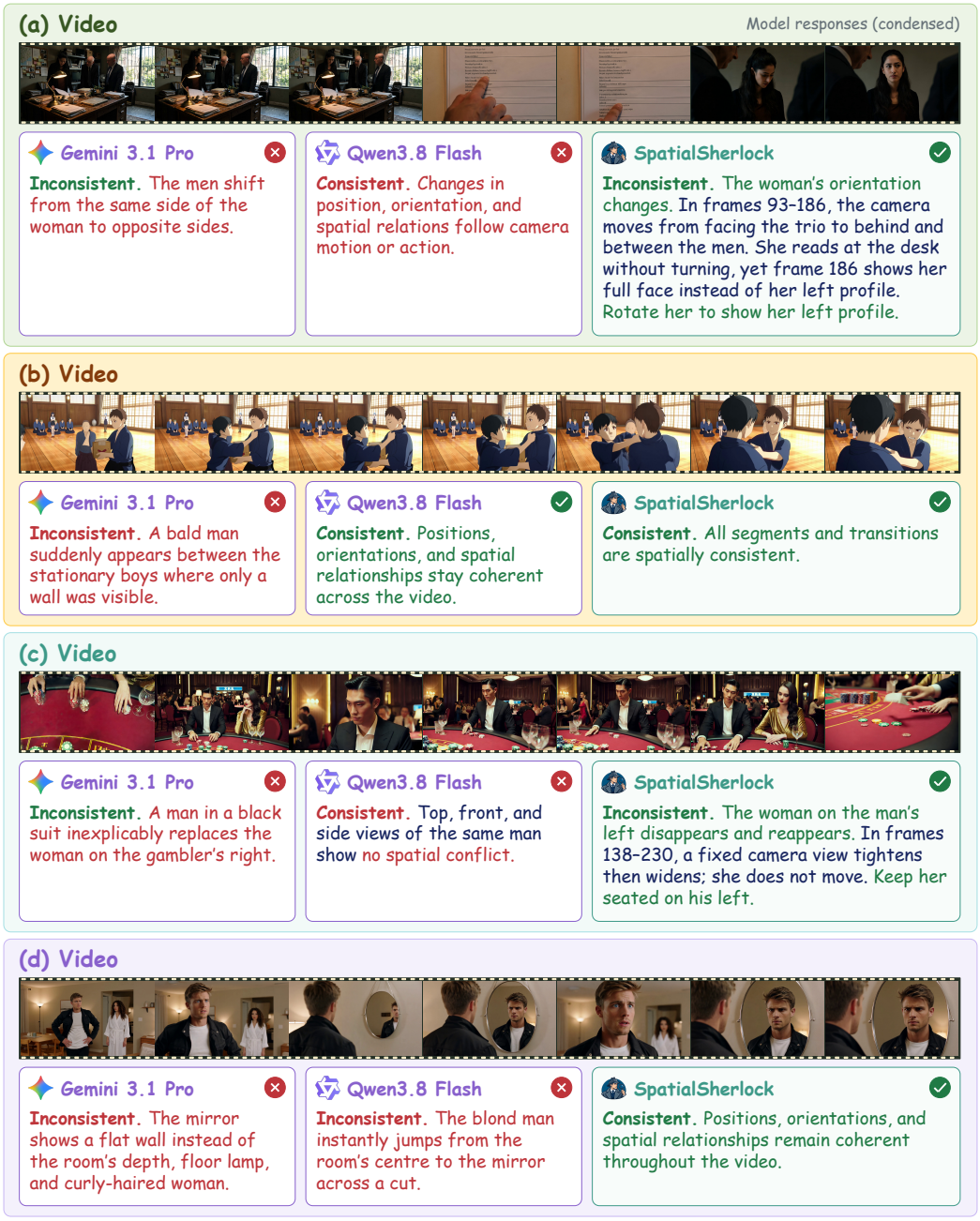

Figure 4: **Qualitative examples across four tasks.** (a,c) contain spatial inconsistencies; (b,d) are spatially consistent. Answers are condensed from the output JSON `response` field (Appendix B). Green/red text marks correct/incorrect claims; check/cross icons indicate whether the diagnosis, including its explanation, is correct.

REFERENCES

Gilad Baruch, Zhuoyuan Chen, Afshin Dehghan, Tal Dimry, Yuri Feigin, Peter Fu, Thomas Gebauer, Brandon Joffe, Daniel Kurz, Arik Schwartz, and Elad Shulman. ARKitScenes - a diverse real-world dataset for 3D indoor scene understanding using mobile RGB-D data. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*, 2021. URL https://openreview.net/forum?id=tjZjv_qh_CE.

BytePlus. Fortal Studio scales | AI-powered short-drama creation, 2026. URL https://www.byteplus.com/en/case-study/Fortal%20Studio. Accessed September 25, 2026.

Zhongang Cai, Ruisi Wang, Chenyang Gu, Fanyi Pu, Junxiang Xu, Yubo Wang, Wanqi Yin, Zhitao Yang, Chen Wei, Tongxi Zhou, Qingping Sun, Hui En Pang, Jiaqi Li, Oscar Qian, Zhiqian Lin, Xuanke Shi, Kewang Deng, Xiaoyang Han, Zukai Chen, Xiangyu Fan, Hanming Deng, Lewei Lu, Liang Pan, Bo Li, Ziwei Liu, Quan Wang, Dahua Lin, and Lei Yang. Scaling spatial intelligence with multimodal foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7879–7890, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Cai_Scaling_Spatial_Intelligence_with_Multimodal_Foundation_Models_CVPR_2026_paper.html.

Boyuan Chen, Zhuo Xu, Sean Kirmani, Brian Ichter, Danny Driess, Pete Florence, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. SpatialVLM: Endowing vision-language models with spatial reasoning capabilities. *arXiv preprint arXiv:2401.12168*, 2024. URL https://arxiv.org/abs/2401.12168.

An-Chieh Cheng, Hongxu Yin, Yang Fu, Qiushan Guo, Ruihan Yang, Jan Kautz, Xiaolong Wang, and Sifei Liu. SpatialRGPT: Grounded spatial reasoning in vision language models. *arXiv preprint arXiv:2406.01584*, 2024. URL https://arxiv.org/abs/2406.01584.

Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. ScanNet: Richly-annotated 3D reconstructions of indoor scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5828–5839, 2017. URL https://openaccess.thecvf.com/content_cvpr_2017/html/Dai_ScanNet_Richly-Annotated_3D_CVPR_2017_paper.html.

Weijia Dou, Wenzhao Zheng, Weiliang Chen, Yu Zheng, Jie Zhou, and Jiwen Lu. Measuring 3D spatial geometric consistency in dynamic video generation. *arXiv preprint arXiv:2603.19048*, 2026. URL https://arxiv.org/abs/2603.19048.

Mengfei Du, Binhao Wu, Zejun Li, Xuanjing Huang, and Zhongyu Wei. EmbSpatial-Bench: Benchmarking spatial understanding for embodied tasks with large vision-language models. *arXiv preprint arXiv:2406.05756*, 2024. URL https://arxiv.org/abs/2406.05756.

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. In *Advances in Neural Information Processing Systems*, 2025. URL https://arxiv.org/abs/2503.21776.

Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A. Smith, Wei-Chiu Ma, and Ranjay Krishna. BLINK: Multimodal large language models can see but not perceive. *arXiv preprint arXiv:2404.12390*, 2024. URL https://arxiv.org/abs/2404.12390.

Xingyu Fu, Siyi Liu, Yinuo Xu, Pan Lu, Guangqiuse Hu, Tianbo Yang, Taran Anantasagar, Christopher Shen, Yikai Mao, Yuanzhe Liu, Keyush Shah, Chung Un Lee, Yejin Choi, James Zou, Dan Roth, and Chris Callison-Burch. Learning human-perceived fakeness in AI-generated videos via multimodal LLMs. *arXiv preprint arXiv:2509.22646*, 2025. URL https://arxiv.org/abs/2509.22646.

Gemini Robotics Team et al. Gemini Robotics: Bringing AI into the physical world. *arXiv preprint arXiv:2503.20020*, 2025. URL https://arxiv.org/abs/2503.20020.

Ting Huang, Zhenyu Zhang, Wenyuan Huang, Jian Yang, and Hao Tang. ConsiSpace: Learning geometric consistency matters for video spatial reasoning. *arXiv preprint arXiv:2607.17599*, 2026. URL https://arxiv.org/abs/2607.17599.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. VBench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. URL https://arxiv.org/abs/2311.17982.

Mukul Khanna, Yongsen Mao, Hanxiao Jiang, Sanjay Haresh, Brennan Shacklett, Dhruv Batra, Alexander Clegg, Eric Undersander, Angel X. Chang, and Manolis Savva. Habitat Synthetic Scenes Dataset (HSSD-200): An analysis of 3D scene scale and realism tradeoffs for ObjectGoal navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16384–16393, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Khanna_Habitat_Synthetic_Scenes_Dataset_HSSD-200_An_Analysis_of_3D_Scene_CVPR_2024_paper.html.

Dingming Li, Hongxing Li, Zixuan Wang, Yuchen Yan, Hang Zhang, Siqi Chen, Guiyang Hou, Shengpei Jiang, Wenqi Zhang, Yongliang Shen, Weiming Lu, and Yueting Zhuang. ViewSpatial-Bench: Evaluating multi-perspective spatial localization in vision-language models. *arXiv preprint arXiv:2505.21500*, 2025. URL https://arxiv.org/abs/2505.21500.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *Computer Vision – ECCV 2014*, pp. 740–755, 2014. doi: 10.1007/978-3-319-10602-1_48.

Wufei Ma, Haoyu Chen, Guofeng Zhang, Yu-Cheng Chou, Jieneng Chen, Celso de Melo, and Alan Yuille. 3DSRBench: A comprehensive 3D spatial reasoning benchmark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6924–6934, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Ma_3DSRBench_A_Comprehensive_3D_Spatial_Reasoning_Benchmark_ICCV_2025_paper.html.

Jiahao Meng, Xiangtai Li, Haochen Wang, Yue Tan, Tao Zhang, Lingdong Kong, Yunhai Tong, Anran Wang, Zhiyang Teng, Yujing Wang, and Zhuochen Wang. Open-o3-Video: Grounded video reasoning with explicit spatio-temporal evidence. *arXiv preprint arXiv:2510.20579*, 2025. URL https://arxiv.org/abs/2510.20579.

Claire Mo. HappyHorse gets stronger motion expressiveness, higher generation consistency, and enhanced visual quality. Alibaba Cloud, June 2026. URL https://www.alibabacloud.com/blog/603293. Accessed September 25, 2026.

Kun Ouyang, Yuanxin Liu, Haoning Wu, Yi Liu, Hao Zhou, Jie Zhou, Fandong Meng, and Xu Sun. SpaceR: Reinforcing MLLMs in video spatial reasoning. *arXiv preprint arXiv:2504.01805*, 2025. URL https://arxiv.org/abs/2504.01805.

Qwen Team. Qwen3.8-Max, 2026. URL https://www.qwencloud.com/models/qwen3.8-max. Official model documentation; accessed September 10, 2026.

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. URL https://arxiv.org/abs/2402.03300.

Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, Ziteng Wang, Rob Fergus, Yann LeCun, and Saining Xie. Cambrian-1: A fully open, vision-centric exploration of multimodal LLMs. *arXiv preprint arXiv:2406.16860*, 2024. URL https://arxiv.org/abs/2406.16860.

Qineng Wang, Baiqiao Yin, Pingyue Zhang, Jianshu Zhang, Kangrui Wang, Zihan Wang, Jieyu Zhang, Keshigeyan Chandrasegaran, Han Liu, Ranjay Krishna, Saining Xie, Jiajun Wu, Li Fei-Fei, and Manling Li. MindCube: Spatial mental modeling from limited views. *arXiv preprint arXiv:2506.21458*, 2025a. URL https://arxiv.org/abs/2506.21458.

Zeqing Wang, Keze Wang, and Lei Zhang. PhyDetEx: Detecting and explaining the physical plausibility of T2V models. *arXiv preprint arXiv:2512.01843*, 2025b. URL https://arxiv.org/abs/2512.01843.

Rui Yang, Ziyu Zhu, Yanwei Li, Jingjia Huang, Shen Yan, Siyuan Zhou, Zhe Liu, Xiangtai Li, Shuangye Li, Wenqian Wang, Yi Lin, and Hengshuang Zhao. Visual spatial tuning. In *European Conference on Computer Vision*, 2026a. URL https://arxiv.org/abs/2511.05491.

Shusheng Yang, Jihan Yang, Pinzhi Huang, Ellis Brown, Zihao Yang, Yue Yu, Shengbang Tong, Zihan Zheng, Yifan Xu, Muhan Wang, Rob Fergus, Yann LeCun, Li Fei-Fei, and Saining Xie. Cambrian-S: Towards spatial supersensing in video. In *International Conference on Learning Representations*, 2026b. URL https://openreview.net/forum?id=rBFDvZu6pb.

Sihan Yang, Runsen Xu, Yiman Xie, Sizhe Yang, Mo Li, Jingli Lin, Chenming Zhu, Xiaochen Chen, Haodong Duan, Xiangyu Yue, Dahua Lin, Tai Wang, and Jiangmiao Pang. MMSI-Bench: A benchmark for multi-image spatial intelligence. *arXiv preprint arXiv:2505.23764*, 2025. URL https://arxiv.org/abs/2505.23764.

Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. ScanNet++: A high-fidelity dataset of 3D indoor scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. URL https://scannetpp.mlsg.cit.tum.de/scannetpp/.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, et al. DAPO: An open-source LLM reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*, 2025. URL https://arxiv.org/abs/2503.14476.

Jiahui Zhang, Yurui Chen, Yanpeng Zhou, Yueming Xu, Ze Huang, Jilin Mei, Junhui Chen, Yu-Jie Yuan, Xinyue Cai, Guowei Huang, Xingyue Quan, Hang Xu, and Li Zhang. From flatland to space: Teaching vision-language models to perceive and reason in 3D. *arXiv preprint arXiv:2503.22976*, 2025. URL https://arxiv.org/abs/2503.22976.

Qin Zhang, Peiyu Jing, Hong-Xing Yu, Fangqiang Ding, Fan Nie, Weimin Wang, Yilun Du, James Zou, Jiajun Wu, and Bing Shuai. Physion-Eval: Evaluating physical realism in generated video via human reasoning. *arXiv preprint arXiv:2603.19607*, 2026. URL https://arxiv.org/abs/2603.19607.

Dian Zheng, Ziqi Huang, Hongbo Liu, Kai Zou, Yinan He, Fan Zhang, Lulu Gu, Yuanhan Zhang, Jingwen He, Wei-Shi Zheng, Yu Qiao, and Ziwei Liu. VBench-2.0: Advancing video generation benchmark suite for intrinsic faithfulness. *arXiv preprint arXiv:2503.21755*, 2025. URL https://arxiv.org/abs/2503.21755.

Wentao Zhu, Yufang Huang, Xiufeng Xie, Wenxian Liu, Jincan Deng, Debing Zhang, Zhangyang Wang, and Ji Liu. AutoShot: A short video dataset and state-of-the-art shot boundary detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 2238–2247, June 2023. URL https://openaccess.thecvf.com/content/CVPR2023W/NAS/html/Zhu_AutoShot_A_Short_Video_Dataset_and_State-of-the-Art_Shot_Boundary_Detection_CVPRW_2023_paper.html.

A FULL-VIDEO REASONING FOR THE TEASER EXAMPLE

The six steps below reproduce SpatialSherlock's original reasoning and final result for the complete video in Figure 1, using the full local-diagnosis, conditional-verification, and video-aggregation pipeline. Frame numbers are one-based. The decisive frames, 243 and 244, are consecutive in the source video: no intervening footage has been omitted between them. Thus, Step 5 refers to a directly adjacent transition, after the spatial comparison in Steps 1–4, rather than a gap between sparsely sampled frames.

Step 1: I compare the two frames. In the first frame, the man in the red shirt is in the left foreground, seen mostly from behind, with his body angled away and his backpack strap across his shoulder.

Step 2: In the second frame, the same red-shirted man is standing more fully visible on the left, turned toward the seated man in the floral shirt. His position relative to the camera, the table, and the seated man has changed.

Step 3: The change is not just a minor pose adjustment or a simple camera pan. His location in the scene and his orientation shift in a way that changes the spatial relationship between him and the people around him.

Step 4: Therefore the frames are not spatially consistent.

Step 5: There is no continuous footage between these two frames to explain the spatial change. The segment is therefore spatially inconsistent.

Step 6: The video contains a spatial inconsistency at frames 243 and 244. The key evidence is the change in the relative position and distance between the red-shirted man and the table.

B STRUCTURED VIDEO OUTPUT

The complete video output is a JSON object with five top-level fields: `response`, `summary`, `answer`, `video_answer`, and `bboxes`. `response` contains the complete natural-language answer, including the judgment, explanation, supporting video evidence, and any required correction. `summary` contains a concise diagnostic explanation used to compute the language reward R_{lang}. `answer` selects the consistency option, whereas `video_answer` selects the temporal-evidence option from the candidates provided with the input. Option letters follow the input question. `bboxes` is a list of region–correction pairs: each item has a `bbox` coordinate array and its associated `correction`. Coordinates follow $[x_1, y_1, x_2, y_2]$ on the normalized $[0, 1000]$ scale of the associated evidence frame; an output with no affected regions uses an empty list.

JSON example. The following values illustrate the format. In this example, the input defines B as inconsistent and C as the supporting video option; the coordinates are illustrative.

```
{
  "response": "Inconsistent. The object jumps. Move it back.",
  "summary": "The object jumps.",
  "answer": "B",
  "video_answer": "C",
  "bboxes": [
    {
      "bbox": [100, 200, 400, 600],
      "correction": "move (100, 450)"
    }
  ]
}
```

The correction couples an action keyword with its parameters; here, `move` specifies a target point. Figures 1 and 4 show condensed `response` text rather than the entire JSON object.

Reward correspondence. The fields correspond to the rewards in Section 4.2 as follows: `answer` supplies C_y, `video_answer` supplies C_t, `bboxes[].bbox` supplies R_{union}, paired `bbox` and

`correction` entries supply R_{match}, and `summary` supplies R_{lang}. The temporal option is evaluated separately with the same exact-match rule as an answer option.

C Data construction and annotation

Spatial-skill data. We use the official releases of HSSD, MS-COCO, ScanNet, ScanNet++, and ARKitScenes listed in Section 3.1. Qwen3.8-Max accesses their images and associated annotations or calibration information to construct questions, answers, and reasoning for object localization, within- and cross-image relative positions, and camera motion. The process uses source geometry and camera parameters where available. Qwen3.8-Max filters the mixed pool, leaving approximately 200,000 retained examples.

Video sources. The shot-by-shot scripts come from actual short-drama and advertising production, grounding the videos in practical narrative and production requirements. More than 20,000 videos of 10–30 s were generated with multiple versions from the Seedance, HappyHorse, MiniMax, Kling, Veo, Vidu, and Grok families, among others. AutoShot segmentation yields approximately 200,000 candidate intervals, with adjacent boundary pairs counted separately.

First screening: visual content. Before annotation, Qwen3.8-Max removes pairs dominated by blur, ghosting, extensive empty shots, or background-only imagery. The corresponding shots are trimmed so that retained frame-pair and video examples contain informative visual content.

Human annotation. Figure 2 summarizes the annotation workflow. A source video stores frames f_t and timestamps τ_t as $V = ((f_t, \tau_t))_{t=1}^{T}$; a local pair is $E = ((f_a, \tau_a); (f_b, \tau_b))$. Annotators first label each retained endpoint or boundary pair with a judgment, discrepancy regions, required corrections, and an explanation. For intervals, they then inspect intervening events to verify or revise the local diagnosis and identify temporal evidence. Across cuts, judgments use scene identity and spatial continuity; corrections to missing content require evidence that it should be visible. The pool contains approximately 120,000 local and interval-level records.

Second screening and reasoning construction. Qwen3.8-Max receives the visual input, task, reference judgment, and available human explanations, region–correction pairs, and temporal annotations. It constructs reasoning and answers while flagging suspected annotation errors for human review. The reviewed pool yields 92,000 task examples: 40,000 Judge, 15,000 Explain, 12,000 Diagnose, and 25,000 Video examples. These are task-example counts; a shared annotation can support multiple tasks.

Teacher prompt.

> You are constructing spatial-diagnosis supervision from video evidence and human annotations. Read the task, images or video, reference judgment, explanation, annotated regions and corrections, and temporal evidence when supplied. First check whether the annotations agree with the visible content. Flag any conflict for human review and explain the conflicting observation; do not invent a rationale to force agreement. For a supported example, write a stepwise explanation linking object positions and spatial relationships to the judgment. Distinguish camera motion, object motion, occlusion, and scene transitions. Preserve the supplied boxes, correction actions, and direction parameters. For Video, use intervening events and the supplied temporal candidates to identify the evidence supporting the interval judgment. For adjacent frames, state that no frames intervene. Return the task answer and its explanation, using only evidence within the supplied visual input.

Data partitions. We partition the diagnostic data by source video, keeping all derived intervals, frame pairs, and task examples in the same partition. The Judge, Explain, and Video test sets are sampled randomly from their respective held-out pools. For Diagnose, annotators additionally flag challenging cases during annotation; we randomly select 300 frame pairs from these held-out cases. Sampling for each task follows the class counts in Appendix F.

D TRAINING AND VIDEO INFERENCE

Training configuration. We initialize from the public Qwen3.5-9B checkpoint and freeze its vision transformer throughout SFT and RL. SFT uses the standard verl configuration with a learning rate of 10^{-5}, mixing spatial-skill, Diagnose, and Explain examples as in Figure 3. RL starts from the SFT checkpoint and follows the verl DAPO recipe for joint training on Judge, Explain, Diagnose, and Video. The regional reward coefficients are $(\alpha, \beta, \gamma, \delta) = (0.1, 0.25, 0.25, 0.4)$; other branches are specified in Appendix E.

For each input x, the rollout policy $\pi_{\theta_{\text{old}}}$ samples G responses $\{a^{(g)}\}_{g=1}^{G}$ with rewards r_g. The advantages and token probability ratios in Equation 2 are

$$A_g = \frac{r_g - \bar{r}}{s_r + \epsilon}, \qquad \bar{r} = \frac{1}{G}\sum_{g=1}^{G} r_g, \qquad \rho_{g,t}(\theta) = \frac{\pi_\theta(a_t^{(g)} \mid x, a_{<t}^{(g)})}{\pi_{\theta_{\text{old}}}(a_t^{(g)} \mid x, a_{<t}^{(g)})}, \tag{9}$$

where s_r is the within-group sample standard deviation and $\epsilon > 0$ stabilizes normalization.

Video inference. AutoShot uses its official configuration; video inputs are sampled to 24 frames. The model first diagnoses endpoint and boundary pairs, then conditions on the local judgment and explanation when checking available intervening frames. Adjacent pairs with no intervening frames retain their local diagnoses. The 78.7% Video result evaluates source-video conclusions after the full local-diagnosis, conditional-verification, and video-aggregation pipeline.

Conditional-verification prompt.

> Review the local diagnosis using the supplied video frames in temporal order. You receive the endpoint frames, the first-stage judgment and explanation, and sampled intervening frames with temporal indices. Identify the spatial claim in the local diagnosis and check whether observed camera motion, object motion, or occlusion explains the endpoint difference. Retain the diagnosis if the evidence supports the discrepancy; revise it if intervening events resolve it. Ground each explanation in visible events, and distinguish adjacent frames from gaps in sampling. Return the interval judgment, a concise explanation of whether the local diagnosis is retained or revised, and the temporal evidence supporting it.

E REWARD CALCULATION DETAILS

Reward components and temporal selection. Equations 3–8 define the correctness gate, component scores, and joint regional reward in the main text. For nonempty reference regions, the correctness gate multiplies the weighted combination of component scores, with coefficients $(\alpha, \beta, \gamma, \delta) = (0.1, 0.25, 0.25, 0.4)$. For Video, temporal localization is formulated as multiple-choice selection among candidates derived from C_{ij} and B_i (Section 3.2). The `video_answer` option is compared with the reference: $C_t = 1$ for a match and $C_t = 0$ otherwise. Equation 4 combines this score with correctness of `answer`.

Structured training responses. Structured joint diagnosis is scored on its answer, the explanation in `summary`, and region–correction pairs. Appendix B describes these components in the video output. Coordinate-only localization targets use box arrays directly. For joint diagnosis, scoring uses the last valid JSON object in the model output. Answers are matched after normalizing case, surrounding whitespace, and outer parentheses. For answer-only scoring, the answer field can also be recovered from otherwise malformed JSON. Malformed box entries are discarded, and missing or non-string `summary` values receive zero language reward.

Region masks and matching. Coordinates in `bboxes[].bbox` are clamped to $[0, 1000]$ and truncated to integers for 1000×1000 union masks. Empty masks score one if both are empty, and zero otherwise. Before computing R_{match}, all parsed predicted `correction` entries, including unmatched ones, undergo a format check. The score is -1 if any `move` correction lacks exactly one parenthesized tuple of two or three finite numbers, or any correction without `move` contains such a tuple. Otherwise, empty matchings score zero. Matching maximizes geometric IoU without a threshold; the denominator is the actual number of matched pairs.

Correction and direction weights. The keyword weight is defined in Equation 7. Case-insensitive word-prefix matching extracts the six action keywords into $K_i^\star$ and $\hat{K}_j$; directional agreement affects only `move`. A two-component tuple (x, y) specifies an endpoint, giving direction $(x - c_x, c_y - y)$ relative to its box center; three components specify a direction vector directly. With a reference direction, $d_{ij} = \max(0, \cos\theta)$ for equal dimensions. Unequal dimensions use their shared xy projection and $d_{ij} = \max(0, \cos\theta - \kappa)$, with penalty coefficient $\kappa = 0.1$. Zero vectors score zero in these comparisons; an unspecified reference direction gives $d_{ij} = 1$.

Language agreement. Qwen3.5-9B compares the predicted `summary` with the reference `summary` to compute R_{lang}. The four classes underlying p_k, in order, are: complete coverage of material reference facts without errors or unsupported claims; at least one complete fact agrees, with omissions only; at least one complete fact agrees alongside errors or unsupported claims; and no complete material fact agrees. Probabilities are normalized across these classes. Empty `summary` strings score zero; strings identical after case and whitespace normalization score one.

Reasoning–answer compatibility. For designated answer-bearing training branches, a gate H is evaluated after a correct final answer. Qwen3.5-9B classifies which question option the last reasoning sentence supports. Missing reasoning or explicit support for the opposite option sets $H = 0$; matching or unclear support sets $H = 1$. The gated task reward is $R_{\text{train}} = HR$. During training, responses that exceed the configured length allowance receive a separate penalty.

Reward branches. Each example uses a reward branch appropriate to its task and annotations. Consistency-judgment examples use $R = C$ and the reasoning–answer gate. Joint-diagnosis examples also use that gate and the main-text combination for nonempty reference boxes. With empty reference boxes, $R = C$ for empty predictions and $R = \eta C$ otherwise, with $\eta = 0.2$. Video examples use the same joint-diagnosis reward, component weights, empty-region rules, and reasoning–answer gate, with $C = C_y C_t$. The consistency option in `answer` and temporal option in `video_answer` must both exactly match their references for a positive task reward. Coordinate-only localization examples divide the maximum matched IoU sum by $\max(n, m)$ for n reference and m predicted boxes, returning zero if either set is empty. Examples with region, correction, and explanation annotations use $R = \beta_{\text{d}} R_{\text{union}} + \gamma_{\text{d}} R_{\text{match}} + \delta_{\text{d}} R_{\text{lang}}$ with weights $(\beta_{\text{d}}, \gamma_{\text{d}}, \delta_{\text{d}}) = (0.25, 0.25, 0.5)$ and without C, H, or the branch for empty reference regions described above. Auxiliary spatial question-answering examples use answer correctness without the reasoning–answer gate. These rewards supervise the available annotations; the diagnostic evaluation separately assesses response correctness as defined in Appendix F.

F Diagnostic evaluation

We compare SpatialSherlock with Gemini 3.1 Pro, Gemini 3.5 Flash, Qwen3.8 Flash, and Qwen3.8-Max on four diagnostic tasks. Judge contains 400 frame pairs: 200 positive (consistent) and 200 negative (inconsistent) examples. Explain and Diagnose each contain 300 frame pairs; Video contains 300 complete source videos. Each 300-example test set contains 100 positive (consistent) and 200 negative (inconsistent) examples. These held-out evaluation counts are distinct from the task-pool sizes in Section 3.2.

Each response receives one point if correct and zero otherwise. Judge reports per-class accuracy and their arithmetic mean. Human evaluators check every explanation against the ground truth: Explain requires a matching explanation, Diagnose requires both the judgment and explanation, and Video additionally requires supporting temporal evidence. Video is scored on the final source-video response after local diagnosis, conditional verification, and video aggregation.

G External benchmark evaluation

All external evaluations follow the benchmarks' official configurations and scoring protocols. SpatialSherlock and its Qwen3.5-9B backbone are evaluated zero-shot. Table 2 reports CV-Bench overall, BLINK validation, and MindCube TinyBench. The comparison models are SenseNova-SI-1.1-Qwen3-VL-8B, VST-7B-RL, and Cambrian-S-7B. The star marks SN-SI-1.1's use of MindCube-10K training data before evaluation on the disjoint TinyBench subset.

H ABLATION STUDIES

Except for the factor under study, all comparisons use the main experiment's training configuration, inference settings, test sets, and evaluation criteria.

H.1 SPATIAL-SKILL AND DIAGNOSTIC DATA

We compare SFT on diagnostic data alone and on a mixture of spatial-skill and diagnostic data from the same backbone under a matched token budget (Table 3). Diagnostic SFT uses Diagnose and Explain, as in the main training pipeline. We apply identical joint RL to the diagnostic-only and mixed SFT checkpoints to test whether the skill-data contribution persists after RL. Both RL runs retain all four diagnostic tasks; data removal applies only to SFT. ViewSpatial-Bench complements diagnostic accuracy with an evaluation of reasoning across views. We also evaluate the spatial-skills-only SFT checkpoint on ViewSpatial-Bench; its diagnostic tasks are not evaluated. Video is not evaluated at the SFT stage because this stage excludes Video training data.

Table 3: Data-source ablation (%). Judge reports the class mean; "/" denotes a task that was not evaluated.

SFT data	Video	Diagnose	Judge	Explain	ViewSpatial
SFT checkpoints					
Spatial skills only	/	/	/	/	46.6
Diagnostic data only	/	60.3	66.0	52.7	49.0
Both sources	/	66.7	72.0	73.3	51.8
After identical joint RL					
Diagnostic data only	75.0	74.7	80.5	71.3	54.4
Both sources	78.7	81.3	83.3	84.0	57.2

H.2 TRAINING STAGES

We compare mixed-data SFT, SFT + continued SFT, and SFT + joint RL (Table 4). The initial SFT checkpoint is evaluated on Diagnose, Judge, and Explain; Video is evaluated after continued SFT or joint RL. To separate RL from the newly introduced Video data, a continued-SFT control starts from the same SFT checkpoint. It uses the same four-task training pool, task mixture, and reference annotations as RL, with comparable compute including rollout and reward evaluation. The continued-SFT objective uses the main SFT loss to supervise the reference reasoning and task outputs for all four tasks.

Table 4: Training stages (accuracy, %); metrics follow Table 1. Judge reports the class mean. "/" denotes a task that was not evaluated.

Training stage	Video	Diagnose	Judge	Explain
SFT	/	66.7	72.0	73.3
SFT + continued SFT	60.3	70.7	78.0	76.3
SFT + RL (SpatialSherlock)	78.7	81.3	83.3	84.0

Shared configuration and evaluation. Training and inference follow Appendix D, and reward branches follow Appendix E, subject to the ablation changes specified above. All other settings, including data sampling, token and compute budgets, optimization, and random seeds, follow the main experiment. Diagnostic comparisons use its source-video partitions, test examples, prompts, decoding settings, and human scoring criteria (Appendix F). Video accuracy uses the same source-video scoring and aggregation rules. External benchmarks follow Appendix G.