MoRe-UAV: A Large-Scale Benchmark for Motion-Aware Visual Grounding in UAV Videos

Zhipeng Zhang
School of Computer Science, Northwestern Polytechnical University
Xi'an, China

Yiheng Zhang
School of Computer Science, Northwestern Polytechnical University
Xi'an, China

Wei Suo
School of Computer Science, Northwestern Polytechnical University
Xi'an, China

Le Liu
School of Computer Science, Northwestern Polytechnical University
Xi'an, China

Ji Wang
School of Computer Science, Northwestern Polytechnical University
Xi'an, China

Peng Wang
School of Computer Science, Northwestern Polytechnical University
Xi'an, China

This supplementary material expands the dataset, method, and experimental descriptions that are abbreviated in the main paper. We focus on the benchmark assets released with MoRe-UAV, the quality-control pipeline behind the annotations, additional implementation notes for the reference baseline, and qualitative analyses that further illustrate the strengths and remaining limitations of the task.

1 Dataset Details

1.1 Benchmark Summary

MoRe-UAV is designed for motion-aware visual grounding in real UAV videos, where both the target and the camera may move simultaneously. In contrast to image grounding or video grounding with relatively stable viewpoints, the benchmark emphasizes two coupled challenges: motion-centric target disambiguation and strong cross-view appearance change caused by UAV ego-motion. This design choice is reflected not only in the language annotations but also in the trajectory statistics summarized in Table 1. The released expressions are intentionally biased toward motion-bearing descriptions rather than short static phrases, and the trajectories exhibit substantial scale variation and direction changes over time.

1.2 Released Benchmark Assets

The public benchmark release excludes raw UAV recordings and only contains anonymized, privacy-screened benchmark assets needed for research use and evaluation. At the sample level, each retained clip is paired with exactly one target object and exactly one verified motion-centric referring expression. In addition to the core video-expression pair, the release package contains:

- processed RGB frames for each released sample;
- frame-level target annotations across the full clip;
- one verified motion-centric referring expression per clip;
- one verified scene label and one verified object label per clip;
- official train/validation/test split assignments;
- evaluation scripts, baseline code, and release documentation;
- documentation describing the data organization, file structure, and usage protocol.

Table 1: Summary statistics of MoRe-UAV used throughout this supplementary material.

Statistic	Value
Video-expression pairs	22,225
Annotated frames	7,415,622
Verified scene categories	28
Verified target categories	30
Average clip duration	~10 s
Average frames per clip	~333
Average words per expression	~11
Average motion terms per query	4.1
Multi-stage motion expressions	39.4%
Relative-motion expressions	28.6%
Average normalized trajectory length	0.71
Average major direction changes per clip	1.22
Median box scale span	4.38×
Train / Val / Test	15,432 / 2,571 / 4,222

This organization keeps the benchmark easy to use for both model training and manual auditing. In particular, the frame-level annotations allow evaluation over full video sequences rather than only over sparse key frames, which is important in UAV videos where severe viewpoint changes can occur gradually across time.

1.3 Video and Acquisition Protocol

These recordings were acquired specifically for the project by the authors together with project-affiliated data collectors, all under a unified acquisition and release protocol. In our terminology, a *source video* denotes a continuous UAV recording obtained during a single collection flight or flight segment. The benchmark samples used for training and evaluation are then created by extracting one or more target-centric clips from these longer source videos. This distinction is important because it makes the provenance of each sample explicit and also underlies the split strategy used later in the benchmark.

Most source videos were captured using DJI consumer-grade UAV platforms equipped with stabilized RGB cameras, while a smaller portion of the footage was collected using other consumer

UAV platforms of comparable capability under the same acquisition protocol. The recordings were obtained during dedicated outdoor flight missions in legally permitted areas, and all participating collectors operated within the same project workflow, privacy requirements, and subsequent data review process. The collection environments cover representative civilian aerial scenes, including urban traffic areas, campuses, mountains, parks, plazas, and outdoor activity sites.

The flight patterns were chosen to preserve realistic UAV motion rather than to simulate a static surveillance camera. During data collection, the UAV actively follows or observes moving targets from varying altitudes, distances, and viewing angles. As a result, the source footage naturally contains ego-motion, target motion, scale variation, and viewpoint transitions from oblique to near top-down views. These properties are central to the motivation of MoRe-UAV and differentiate it from benchmarks dominated by stable viewpoints.

In the processed benchmark release, each retained sample is organized by a unique sample identifier and is associated with processed RGB frames, the corresponding frame-level bounding boxes, and the verified motion-centric expression. These released assets are aligned through consistent sample identity and frame indexing, which makes the benchmark straightforward to use for training, evaluation, and manual inspection. Regardless of who recorded the original source video, every sample entering the benchmark is required to pass the same anonymization, privacy screening, and manual release review before annotation and publication.

1.4 Annotation Pipeline and Quality Control

The dataset is built through a privacy-aware human-in-the-loop pipeline. Starting from the project-collected source videos described above, annotators identify target-centric segments that are suitable for motion-aware grounding and exhibit meaningful viewpoint change. This source-video-first organization allows multiple clips to be extracted from the same recording when they correspond to distinct targets or motion events, while still preserving the provenance of each clip for later split assignment and quality tracking.

Trajectory annotation is performed clip by clip. Annotators first select a target-centric segment and manually mark an initial bounding box for a single target. A tracker is then used only to propagate preliminary box proposals, after which every retained clip is manually inspected and corrected at the frame level. Clips are kept only when the target remains visible for at least 90% of the clip; otherwise, the clip is truncated or discarded. Clips with unstable trajectories that remain unreliable after manual correction are also removed.

For language annotation, the pipeline first creates a seed set of 1,000 clips with manually written motion-aware descriptions, scene labels, and object labels. A locally deployed Qwen2.5-VL assistant, adapted with prompt tuning, is then used to generate draft expressions and draft labels for the remaining candidates. These model outputs are never released directly. Instead, every candidate sample is sent to at least two trained annotators for human verification. Reviewers must confirm that the expression explicitly relies on motion cues, that the clip exhibits observable UAV viewpoint change, that the target can be unambiguously localized by a human, and that the frame-level boxes remain consistent with the referred target throughout the sequence. If a reviewer edits the draft expression, the edited version is treated as a new sample version and must be re-approved independently.

The final quality control stage combines per-sample dual verification with a daily audit. At least two annotators must approve the same final version before a sample is accepted. In addition, 10% of each annotator's daily workload is randomly spot-checked, and the daily batch is accepted only when the audit pass rate exceeds 95%. Starting from 40,320 candidate clip-expression pairs, this filtering pipeline retains 22,225 final samples, corresponding to an acceptance rate of 55.1%.

1.5 Benchmark Organization and Split Strategy

Following the task definition in the main paper, a MoRe-UAV sample can be written as

$$\mathcal{S} = (V, Q, B, y^{\text{scene}}, y^{\text{obj}}), \tag{1}$$

where $V = \{I_t\}_{t=1}^{T}$ is a UAV video clip, Q is a motion-centric referring expression, $B = \{b_t\}_{t=1}^{T}$ is the frame-level grounding sequence, and y^{scene} and y^{obj} denote the verified scene and target labels. For each frame t, b_t is a bounding box when the target is visible and an empty label otherwise. This full-sequence formulation is essential for MoRe-UAV, because the target may remain identifiable only when the model reasons jointly over motion, temporal context, and viewpoint changes across the whole clip.

The benchmark organization is therefore centered on full clip-level supervision rather than sparse key-frame annotation. Each released sample contains the motion-centric query together with the corresponding frame-level target trajectory across the complete clip. The verified scene label and object label serve as additional metadata for analysis and split construction, while the natural-language query remains the primary supervision signal for grounding.

At the dataset level, these semantic labels are consolidated into 28 scene categories and 30 target categories after verification. They are useful for summarizing benchmark diversity and for checking whether the train/validation/test partitions remain balanced across object types. They are not meant to simplify the task into category classification; instead, they complement the motion-aware expressions with lightweight structured metadata.

To prevent leakage, all clips from the same source video are assigned to the same split. The final partition contains 15,432 training samples, 2,571 validation samples, and 4,222 test samples. During split construction, the dataset is made roughly proportional across verified object categories while ensuring that every verified object label appears in all three splits. This strategy preserves category diversity while avoiding the optimistic bias that would arise if visually similar clips from the same source video were distributed across multiple splits.

2 Methods Details

This section provides additional design details for the reference baseline introduced in the main paper. The baseline is built on a frozen multimodal large language model and updates only a lightweight set of task-specific parameters. This design isolates task-specific adaptation from backbone scale, makes the reference

system easier to reproduce, and still allows the model to adapt to the motion-centric and cross-view nature of MoRe-UAV.

2.1 Reference Baseline Overview

Our reference baseline is instantiated on top of frozen Qwen2.5-VL and MiniCPM backbones using parameter-efficient tuning. The overall design is motivated by the two dominant failure sources in MoRe-UAV: motion-centric ambiguity in the language query and rapid appearance variation caused by camera motion. To address them, we introduce two lightweight modules: the Motion-aware Prefix Adapter (MPA), which improves motion-sensitive language conditioning, and the Multi-view Alignment Adapter (MVA), which improves cross-frame feature consistency under dynamic viewpoints.

The complete model processes the input video together with a motion-centric query and predicts a frame-level grounding sequence across the clip. Importantly, the model is designed for direct grounding rather than first-frame identification followed by pure temporal propagation. This allows the decoder to keep using both visual evidence and language cues throughout the sequence instead of relying entirely on tracker state after initialization.

2.2 Motion-aware Prefix Adapter

The key observation behind MPA is that motion-centric grounding depends on more than isolated verbs. In many queries, the target is identified by a structured composition of action words, temporal connectors, and relational phrases, such as following, turning, crossing, or moving behind another object. Directly appending a hand-crafted motion prompt can help, but it usually fails to capture which parts of the sentence are most discriminative for the final grounding decision.

Let $\mathbf{L} = \{\mathbf{l}_i\}_{i=1}^{N}$ denote the query token embeddings, and let $\tilde{\mathbf{y}} = \{\tilde{y}_i\}_{i=1}^{N}$ denote weak motion pseudo-labels extracted from motion-bearing verbs and directional phrases. MPA computes motion-aware token weights

$$\alpha_i = \mathrm{softmax}(\mathbf{w}_m^{\top}\mathbf{l}_i), \tag{2}$$

and supervises them with the weak labels through

$$\mathcal{L}_{\mathrm{motion}} = -\sum_{i=1}^{N} \tilde{y}_i \log \alpha_i. \tag{3}$$

The resulting motion-sensitive query representation is

$$\mathbf{z}_m = \sum_{i=1}^{N} \alpha_i \mathbf{l}_i. \tag{4}$$

This aggregated representation is projected into a small set of learned prefix prompts and prepended to the decoder input. In practice, this design gives the model a soft, learnable notion of which query tokens matter most, instead of forcing it to rely only on explicit keyword matching.

An important advantage of MPA is that the weak labels are used only as lightweight training guidance. At inference time, the adapter no longer depends on manual motion prompts; it uses the learned scorer to emphasize the most informative motion-sensitive tokens automatically. This is especially useful for multi-stage expressions in MoRe-UAV, where the grounding signal may be distributed across several parts of the sentence.

2.3 Multi-view Alignment Adapter

MPA addresses the language side of the problem, while MVA addresses the visual side. In UAV videos, the same target can change scale, orientation, and apparent shape rapidly because of UAV ego-motion. A simple frame-to-frame propagation strategy is therefore brittle: once the intermediate state becomes unreliable, drift can accumulate quickly.

Given a visual representation $\mathbf{v}_t$ for frame t and a propagated alignment state $\mathbf{s}_{t-1}$ from adjacent temporal context, MVA computes an aligned feature

$$\mathbf{c}_t = f_a([\mathbf{v}_t; \mathbf{s}_{t-1}]), \tag{5}$$

where $f_a(\cdot)$ is a lightweight alignment and fusion function. Because the propagated context may be harmful under blur, occlusion, or abrupt camera motion, MVA further applies a confidence gate:

$$\mathbf{h}_t = g_t \cdot \mathbf{c}_t + (1 - g_t) \cdot \mathbf{v}_t, \quad g_t = \sigma\big(f_g([\mathbf{v}_t; \mathbf{c}_t])\big), \tag{6}$$

where $f_g(\cdot)$ is a lightweight gating function. The enhanced representation $\mathbf{h}_t$ is then converted into alignment prompts for the decoder.

Conceptually, MVA differs from a naive iterative strategy in two ways. First, it performs alignment in feature space rather than simply copying a previous prediction into the next frame. Second, it allows the model to decide how much temporal context should be trusted at each step. This is particularly important in aerial videos with occlusion, motion blur, and large perspective shifts.

2.4 Relation to Other Baselines

All direct MLLM baselines use the same backbone family and the same parameter-efficient tuning budget as our method, but remove MPA and MVA. This keeps the comparison focused on the value of motion-sensitive query modeling and temporal cross-view alignment rather than on backbone scale alone.

For MLLM+tracking pipelines, the MLLM is used primarily to identify the target in the initial stage, and a tracker is then responsible for propagating the prediction to later frames. This design can be competitive when the target remains visually stable, but it becomes vulnerable when the scene contains severe viewpoint change, prolonged ambiguity, or partial occlusion. In contrast, our reference baseline keeps language-conditioned visual reasoning active throughout the video.

For STVG baselines such as Co-grounding and CG-STVG, we follow the official training protocols when applicable and adapt them to the MoRe-UAV split. These methods provide useful reference points for spatio-temporal grounding, but they are not originally designed for the particularly strong viewpoint dynamics present in UAV videos.

3 Experiment Details

3.1 Evaluation Protocol

We evaluate all methods on the same official train/validation/test split of MoRe-UAV. Following prior video grounding and tracking practice, we report mIoU, Acc@0.5, Norm Precision, and IoU AUC.

Together, these metrics capture both overlap quality and localization precision across the full video sequence.

The evaluation protocol operates on frame-level predictions over the entire clip rather than on a sparse subset of frames. Frames in which the target is not visible are annotated as empty and remain part of the evaluation sequence. This prevents methods from being rewarded for focusing only on easy segments and better reflects the real UAV setting, where a target may become small, partially occluded, or temporarily difficult to localize even inside an otherwise valid clip.

3.2 Compared Settings

We benchmark three method families in the main paper: spatio-temporal video grounding models, direct MLLM baselines, and MLLM+tracking pipelines. All methods are trained or adapted on the same benchmark split. For direct MLLM baselines, we use the same frozen backbone family and the same parameter-efficient tuning budget as our method, but remove the proposed MPA and MVA components. For MLLM+tracking pipelines, the MLLM is used for target identification while the tracker propagates boxes to subsequent frames. For STVG models, we follow the original training protocols whenever they are applicable and then evaluate on the MoRe-UAV split.

This unified setting is important for fair comparison. The direct MLLM and our method differ mainly in whether motion-aware query reweighting and cross-view alignment are introduced. The MLLM+tracking pipelines isolate a different design choice: they test whether a strong initial grounding signal can be maintained by tracker propagation alone. The results in the main paper indicate that this strategy is systematically insufficient once viewpoint change and motion ambiguity accumulate over time.

3.3 Human Evaluation

To contextualize the difficulty of MoRe-UAV, we additionally report human performance on a randomly sampled 20% subset of the test set. Human annotators are allowed to pause and replay videos during evaluation, and their annotations are scored with the same metrics used for model predictions. This protocol measures how well a careful human evaluator can localize the target when both the motion-centric query and the entire clip are available.

The large gap between model performance and human performance highlights that MoRe-UAV is far from saturated. Even though the proposed method substantially improves both the Qwen and MiniCPM baselines, the benchmark still leaves ample room for future research on motion reasoning, cross-view consistency, and robust small-target discrimination under realistic UAV acquisition conditions.

3.4 Additional Notes on Experimental Trends

The ablation results in the main paper reveal two consistent patterns. First, simple motion-word prompting already improves the frozen Qwen baseline, confirming that motion cues are indeed important in this benchmark. However, MPA brings larger gains because it models motion-sensitive composition rather than only inserting isolated keywords. Second, appending previous predictions as a naive temporal hint yields only moderate improvement, while MVA performs better by explicitly aligning temporally adjacent visual representations and adaptively controlling how much temporal context should be trusted.

These trends are consistent with the underlying structure of MoRe-UAV. Many queries require recognizing not only what an object looks like, but also how it moves relative to the scene or to nearby objects. Likewise, many failure cases originate not from the first frame, but from progressive drift caused by viewpoint change and accumulated ambiguity over time.

4 Visualization

4.1 Successful Qualitative Examples

Figure 1 presents several successful qualitative examples of our method. These cases highlight the intended behavior of the reference baseline under diverse UAV viewpoints and motion-centric expressions.

The first case is particularly representative of the strengths of the proposed design. The model not only recognizes the relevant traffic context, but also connects the turning behavior described in the text to the correct bus in the appropriate lane. This suggests that the combination of multimodal scene understanding and motion-aware query modeling is working as intended. The remaining examples show that the model can also remain effective in visually diverse settings, including outdoor natural scenes and fast-moving traffic sequences, where viewpoint and scale may change substantially across frames.

4.2 Representative Failure Cases

Figure 2 summarizes representative failure cases for three model families: direct MLLM grounding, MLLM+tracker pipelines, and our method.

For direct MLLM grounding, we observe a recurring tendency to output overly large bounding boxes in visually uncertain scenes, especially under low light or weak local contrast. In such cases, the model often captures a coarse region that is semantically relevant but spatially imprecise. This behavior is consistent with the strengths and weaknesses of large multimodal models: they can understand the scene broadly, but may still struggle with precise small-object localization when the visual evidence is noisy.

For MLLM+tracker pipelines, the main issue is not only initialization accuracy but also the lack of continued language-grounded understanding in later frames. Once the tracker takes over, prolonged ambiguity, partial occlusion, or scene clutter can cause the prediction to drift away from the referred target. In some cases, the predicted box may even remain fixed at a stale location while the actual target continues to move. This confirms that target propagation alone is insufficient for MoRe-UAV, where the semantic meaning of the query needs to remain active throughout the sequence.

Our method reduces both of the above failure modes by combining the Motion-aware Prefix Adapter with the Multi-view Alignment Adapter, so that motion relations remain emphasized and temporal viewpoint variation is handled more explicitly. Nevertheless, the bottom row of Fig. 2 shows that the problem is not solved completely. When the target is extremely similar to the surrounding environment or to nearby distractors, the model may still fail to distinguish the correct instance precisely. This suggests that future

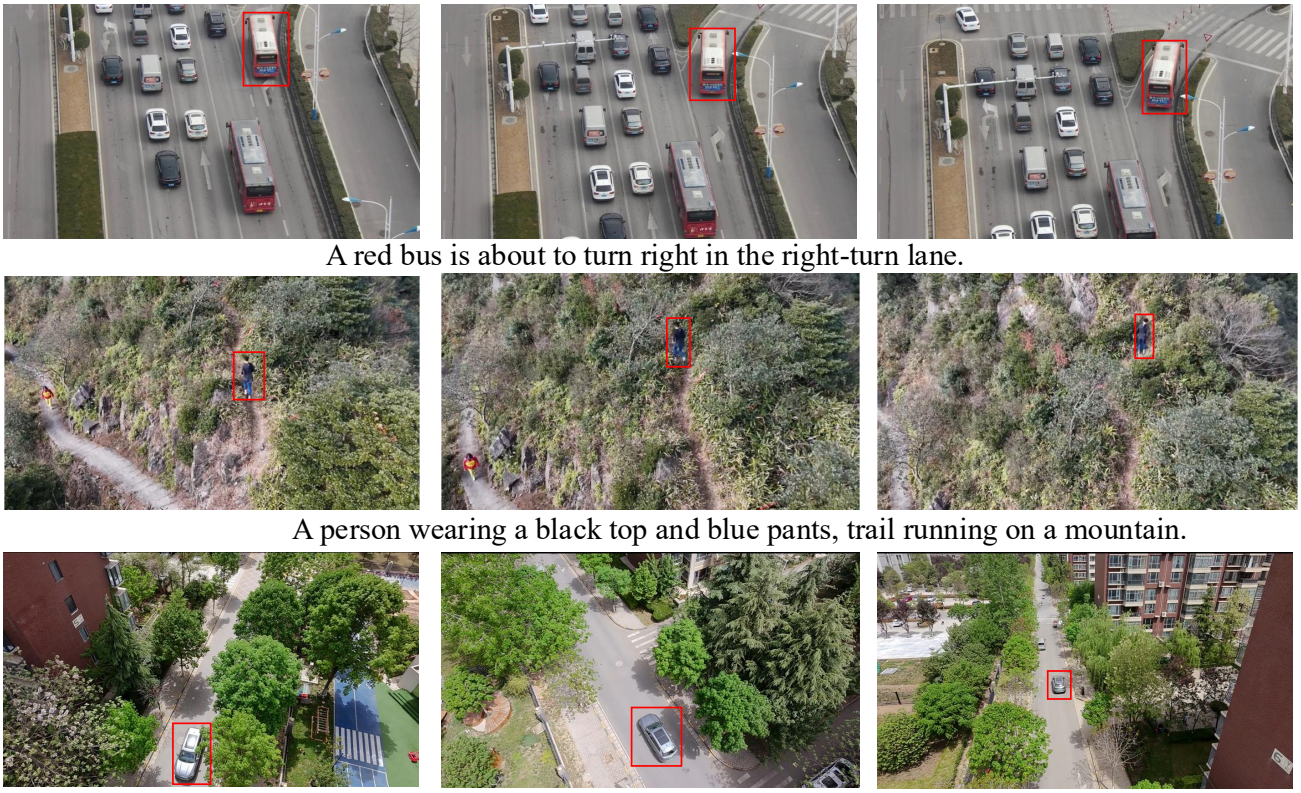

Figure 1: Successful qualitative examples of our method. From left to right, the examples correspond to: a red bus about to turn right in the right-turn lane, a person trail running on a mountain, and a small gray sedan moving straight at high speed. These cases illustrate that the model can jointly leverage scene understanding, target appearance, and motion-centric language cues.

improvements should place even more emphasis on fine-grained target-scene contrast, long-range temporal identity reasoning, and uncertainty-aware localization in difficult aerial scenes.

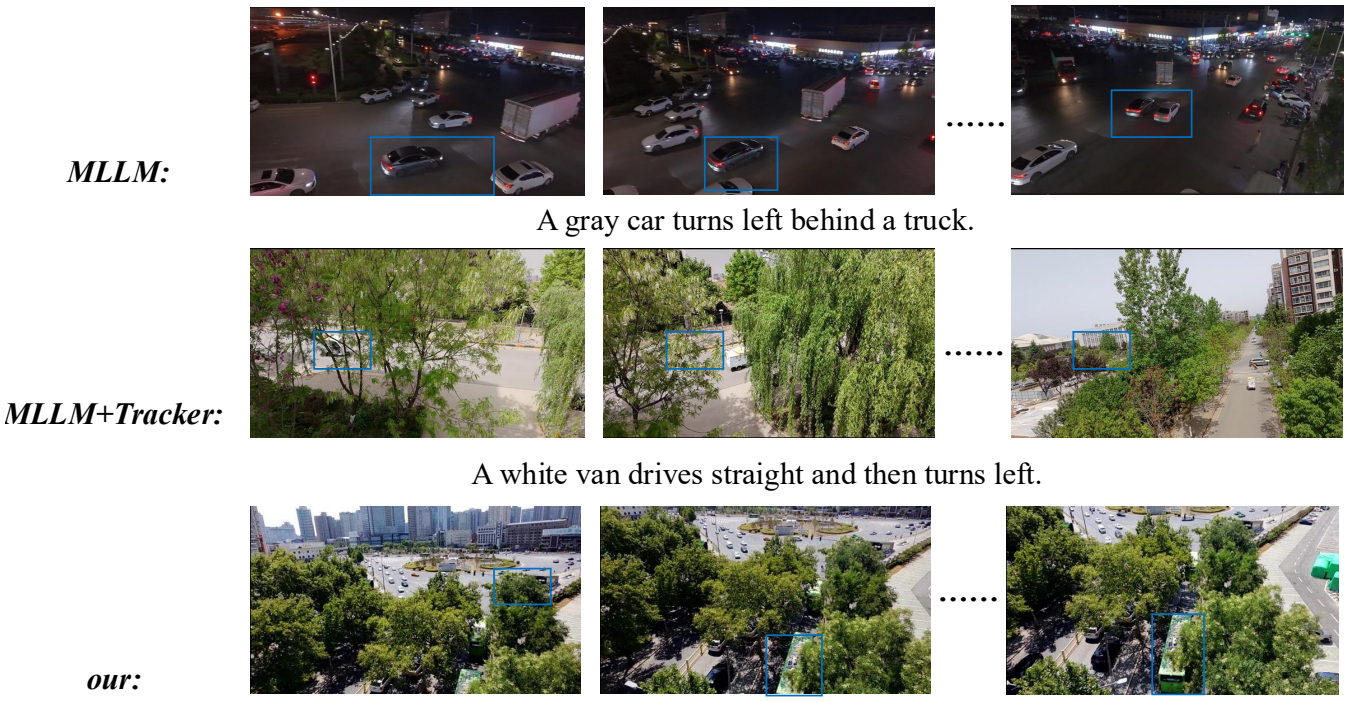

Figure 2: Representative failure cases. Top: direct MLLM grounding with Qwen tends to produce overly large boxes in uncertain conditions such as low-light scenes. Middle: MLLM+tracker may drift, keep predicting a stale location, or become unstable once later frames are ambiguous or partially occluded. Bottom: our method still fails when the target is extremely similar to the surrounding environment and the available discriminative cues are too weak for precise separation.