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

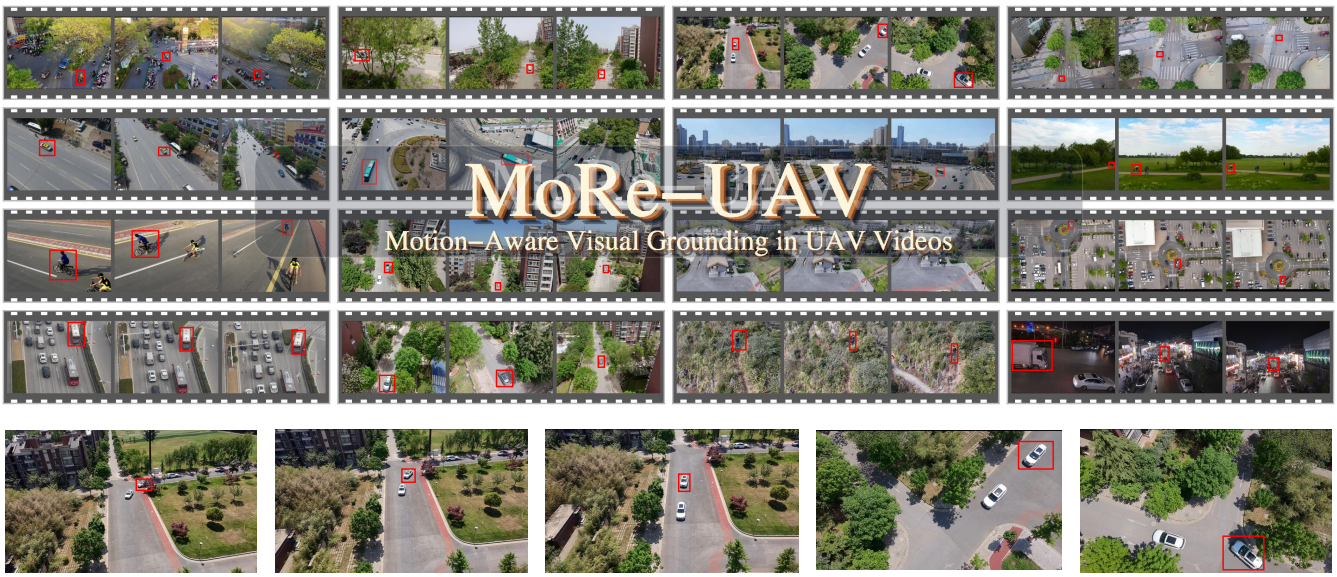


Figure 1: Overview of MoRe-UAV, a large-scale benchmark for motion-aware visual grounding in UAV videos. Top: representative UAV clips with moving targets, camera motion, and large viewpoint changes. Bottom: a typical task example. Unlike conventional UAV grounding that mainly relies on appearance or spatial cues, MoRe-UAV requires reasoning about motion cues under moving viewpoints.

Abstract

UAV visual grounding in real-world applications requires localizing a target referred to by language while both the target and the UAV move. Existing UAV grounding datasets mainly focus on images, while the few video-based benchmarks are still dominated by appearance and spatial cues. As a result, they do not adequately capture two key challenges: motion-centric grounding and drastic cross-view appearance changes caused by UAV ego-motion. To address this gap, we introduce **MoRe-UAV**, a large-scale benchmark for motion-aware visual grounding in UAV videos. MoRe-UAV contains 22,225 video-expression pairs and 7,415,622 annotated frames, covering diverse aerial scenes with moving targets and substantial viewpoint changes. We build the dataset through a scalable human-in-the-loop pipeline for efficient annotation with quality control. We establish an initial benchmark on MoRe-UAV with

Conference'17, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

spatio-temporal video grounding methods, multimodal large language models, and hybrid MLLM+tracking pipelines. We further provide a stronger baseline with a Motion-aware Prefix Adapter and a Multi-view Alignment Adapter to enhance motion reasoning and cross-view alignment. Experiments show that existing methods struggle on MoRe-UAV and remain far below human performance, highlighting substantial room for future research on motion-aware and multi-view grounding in UAV videos. Project page: https://more-uav.github.io/

CCS Concepts

• **Computing methodologies** → **Computer vision**.

Keywords

Visual grounding, UAV Video Dataset.

1 Introduction

Visual grounding [**?** **?** **?**] aims to localize an object referred to by natural language. In UAV applications [**?** **?**] such as search and rescue, aerial surveillance, and autonomous inspection, grounding often needs to operate on video streams rather than isolated images. However, most existing grounding benchmarks [**?** **?** **?**] still emphasize static appearance, spatial attributes, and relatively stable viewpoints [**?**]. This mismatch [**?** **?**] is especially pronounced in UAV scenarios, where the platform moves continuously and the same target can undergo large changes in scale and appearance over time.

We argue that UAV video grounding should be centered on two core challenges. The first is **motion-centric grounding**. In real flights [**?**], the referred target is often distinguished by behaviors, trajectories, or interactions, such as "the white vehicle that follows another car and then turns right." Resolving such queries requires jointly reasoning about motion in both the video and the language description, rather than relying only on static category, color, or position cues. The second is **cross-view appearance change**. UAV ego-motion [**?**] continuously changes altitude, distance, and viewing angle, often shifting from oblique views to top-down observations, so the same object may look substantially different across frames. Existing UAV grounding datasets [**?** **?** **?**] only partially capture this setting: image benchmarks miss temporal motion, while current UAV video benchmarks remain dominated by spatial descriptions and relatively mild viewpoint changes.

In contrast to vision-language tracking [**?** **?** **?** **?**], which mainly emphasizes temporal propagation and association of language-specified targets, our task focuses on grounding a single referred target directly from a motion-centric query in UAV videos with severe viewpoint changes. This shifts the emphasis from target propagation to semantic disambiguation and motion-aware video-language understanding. We therefore study **motion-aware visual grounding in UAV videos**: given a UAV video and a motion-centric query, the goal is to localize the referred target across video frames. Fig. 1 illustrates representative UAV scenarios and a typical task example in MoRe-UAV.

To support this problem, we introduce **MoRe-UAV**, a large-scale benchmark for motion-aware visual grounding in UAV videos, collected from real UAV flights. MoRe-UAV contains 22,225 video-expression pairs and 7,415,622 annotated frames from diverse aerial scenes, with moving targets and substantial viewpoint changes. The dataset is constructed through a scalable human-in-the-loop pipeline with privacy anonymization, manual box correction, human expression review, and consensus verification by at least two trained annotators. On top of this resource, we establish an initial benchmark covering spatio-temporal video grounding methods, direct multimodal large language model baselines, and hybrid MLLM+tracking pipelines. We further provide a stronger baseline with a Motion-aware Prefix Adapter and a Multi-view Alignment Adapter to strengthen motion reasoning and cross-view alignment. The benchmark package has been prepared for academic release via the project page, including processed video frames, frame-level annotations, motion-centric expressions, verified labels, train/val/test splits, evaluation scripts, baseline codes, and documentation with data format. The dataset is intended for non-commercial research use under CC BY-NC 4.0, and the benchmark code is released under the MIT License. To support responsible release, all videos are anonymized before publication, and privacy-sensitive samples are removed during verification.

Our main contributions are summarized as follows:

- We introduce **MoRe-UAV**, a large-scale benchmark for motion-aware visual grounding in UAV videos, with 22,225 video-expression pairs and 7,415,622 annotated frames. The benchmark highlights motion-centric grounding and substantial viewpoint changes caused by UAV ego-motion.
- We establish an initial benchmark on MoRe-UAV covering spatio-temporal video grounding methods, multimodal large language models, and hybrid MLLM+tracking pipelines, and show that existing methods remain far below human performance on this task.
- We provide a stronger reference baseline with a Motion-aware Prefix Adapter and a Multi-view Alignment Adapter to improve motion reasoning and cross-view alignment.

2 Related Work

2.1 Spatio-Temporal Video Grounding

Recent visual grounding research has extended from static images to video settings, including video object grounding (VOG) [**?** **?** **?** **?**] and spatio-temporal video grounding (STVG) [**?** **?** **?** **?** **?**]. Among them, STVG is more closely related to our setting, as it typically focuses on localizing a single referred target across a video segment. Representative benchmarks such as VidSTG [**?**] and HC-STVG [**?**] have substantially advanced research in this direction. However, they are mostly built on fixed camera videos with relatively stable viewpoints, which differ markedly from UAV videos with continuous ego-motion and stronger cross-view variation.

Related efforts have also begun to emphasize motion-aware grounding beyond conventional STVG settings. MASS [**?**] highlights the importance of explicit spatio-temporal reasoning when motion dynamics are central to understanding. However, such efforts are not designed for UAV videos, where persistent ego-motion and drastic viewpoint changes create a distinct grounding challenge.

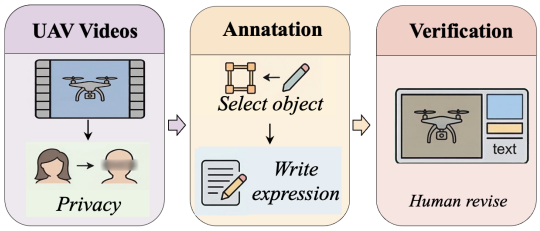


Figure 2: Overview of the annotation pipeline for constructing MoRe-UAV.

2.2 UAV Visual Grounding

With the increasing availability of aerial vision, recent works [? ? ? ? ? ? ?] have begun exploring visual grounding in UAV environments. AerialVG [?] introduced a visual grounding dataset specifically designed for aerial images. These efforts primarily focus on grounding objects in aerial images, extending traditional referring expression comprehension to the UAV domain. These datasets provide valuable benchmarks [? ?] for studying language-guided object localization under aerial viewpoints, but they operate on static images and therefore cannot capture temporal motion patterns.

More recently, UAV-SVG [?] extends grounding to aerial videos, but it is constructed from tracking data [?] and mainly emphasizes appearance-centric and spatial-centric expressions. In contrast, MoRe-UAV focuses on motion-aware grounding in real UAV videos with moving targets and substantial viewpoint changes caused by UAV ego-motion.

2.3 Motion-aware Vision-Language Understanding

Understanding motion described in natural language is an important capability for vision-language models operating in dynamic environments [? ? ?]. MeViS [?] introduces a video object segmentation benchmark that emphasizes motion-centric referring expressions. MotionBench [?] proposes a motion-oriented multimodal benchmark, formulated mainly as video question answering, to evaluate the motion reasoning capabilities of multimodal large language models. Talk2Event [?] extends language grounding to event-camera dynamic scenes. These works highlight the importance of modeling temporal dynamics and motion cues when interpreting language that describes actions or behavioral changes.

However, motion-aware visual grounding in UAV videos remains largely underexplored. In UAV scenarios, motion information plays a crucial role not only in describing target behavior but also in resolving ambiguity among multiple objects in complex scenes. Moreover, UAV ego-motion introduces continuous viewpoint changes, causing significant appearance variations of the same object across frames. These factors make motion-centric grounding substantially more challenging than conventional video grounding tasks. The proposed MoRe-UAV benchmark is designed to address this gap by emphasizing motion-centric grounding and dynamic UAV viewpoints in real aerial videos.

3 MoRe-UAV Dataset

3.1 Task Definition

We study **motion-aware visual grounding in UAV videos**, where the query describes dynamic motion and the referred target moves over time. Given an aerial video $V = \{I_t\}_{t=1}^{T}$ and a motion-centric query Q, the goal is to predict a frame-level grounding sequence $B = \{b_t\}_{t=1}^{T}$, where b_t is a bounding box when the target is visible and empty otherwise. Unlike conventional visual grounding, this task requires jointly reasoning about motion cues in both language and video, while handling continuous viewpoint changes caused by UAV ego-motion.

3.2 Dataset Construction

As shown in Fig. 2, we build MoRe-UAV with a privacy-aware, human-in-the-loop workflow designed around quality-controlled release rather than raw scale. The construction procedure contains five steps.

Step 1: Video data collection. All videos in MoRe-UAV are self-collected through dedicated UAV flight missions using consumer-grade UAV platforms. The recordings cover diverse real-world environments, including urban traffic, campuses, mountains and outdoor activity sites. During collection, the UAV actively follows or observes moving targets from different altitudes, distances, and flight trajectories, producing natural ego-motion, scale changes, and viewpoint transitions from oblique to top-down views. The videos are collected in legally permitted outdoor areas by authorized team members. Detailed metadata on UAV platforms, frame rates, and resolution ranges are provided in the supplementary material.

Step 2: Privacy and ethical review. Our data collection, annotation, and release pipeline is guided by a human-rights-oriented review checklist informed by the Toronto Declaration [?], with particular attention to privacy protection, non-discrimination, and responsible public release. Before annotation and release, all raw UAV videos are anonymized by automatically blurring faces and vehicle license plates[1]. The automatically anonymized videos are then manually screened again, and clips that remain privacy-sensitive or otherwise unsuitable for release are removed. To ensure secure and ethical handling of sensitive UAV footage, all annotation and verification work is conducted by a trained in-house team rather than anonymous crowdsourcing. Annotators receive systematic written instructions on privacy screening and responsible data handling, and their work is compensated according to local institutional and labor requirements. The data collection, annotation, and release procedures were internally reviewed under institutional research and data compliance guidelines. No separate human-subject intervention was involved, and the released data were anonymized and manually screened for privacy-sensitive content prior to publication. The public release excludes raw UAV recordings and contains only anonymized, privacy-screened benchmark data required for research use and evaluation. The release is intended for academic research on motion-aware grounding and video-language understanding, not for identity recognition, surveillance targeting, or other safety-critical decision-making.

[1] https://help.aliyun.com/imm/video-data-anonymization

Table 1: Comparison of representative UAV grounding benchmarks. Existing datasets mainly focus on image-level grounding or UAV video grounding with appearance-centric or spatial-centric expressions. In contrast, MoRe-UAV targets *motion-aware visual grounding in UAV videos*, featuring motion-centric grounding, moving targets, strong viewpoint changes caused by UAV ego-motion, and frame-level target annotations.

Dataset	Type	Source	#Videos / Images	#Frames	Expression Focus	Frame-level Ann.	Viewpoint Change	Motion-aware
AerialVG [?]	Image	VisDrone2019 [?]	5,000	–	Appearance-centric	✗	✗	✗
DVGBench [?]	Image	ERA [?] + VisDrone [?]	2,863	–	Appearance-centric	✗	✗	✗
UAVIT-1M [?]	Image	low-altitude datasets [? ? ?]	789K	–	QA-style	✗	✗	✗
UAV-SVG [?]	Video	CapERA [?] + WebUAV-3M [?]	3,564	2.01M	Spatial-centric	✓	✗	✗
MoRe-UAV (Ours)	**Video**	**Self-collected**	**22,225**	**7.41M**	**Motion-centric**	✓	✓	✓

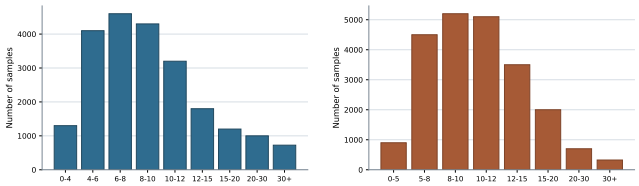
Step 3: Trajectory annotation. From each source video, annotators identify one or more target-centric clips for annotation. For each clip, annotators first select a starting frame and manually annotate an initial bounding box for a single target object. A visual tracking model [?] is then used to propagate preliminary frame-level boxes to the remaining frames. These tracker outputs serve only as annotation proposals rather than final labels. Because UAV videos often involve substantial viewpoint, scale, and appearance changes, every retained clip undergoes manual frame-level inspection and correction to repair drifting or inaccurate boxes. Clips are retained only when the target remains visible for at least 90% of frames; otherwise, the clip is truncated or discarded. Clips whose trajectories remain highly unstable even after manual correction are also removed. Only clips that pass this rigorous manual quality assurance process are included in the final release.

Step 4: Motion-centric expression generation. After obtaining frame-level object trajectories, we create draft motion-centric referring expressions for the moving target. We first manually annotate 1,000 video clips with motion-aware descriptions, scene labels, and object labels to create a seed annotation set. Based on these examples, we adapt a locally deployed Qwen2.5-VL model [?] with prompt tuning [?], enabling it to generate draft motion-aware descriptions conditioned on both the video content and the target trajectory. We simultaneously prompt the model to produce one draft scene label and one draft object label for each clip without restricting the label vocabulary. The adapted Qwen model is used only to provide initial text and label drafts rather than released annotations. Each retained clip is paired with exactly one target object and one draft referring expression for human verification. To prevent data leakage, all clips from the same source video are consistently assigned to the same train/validation/test split.

Step 5: Human verification. The semi-automatic pipeline produces 40,320 candidate video-expression pairs, which are then filtered through a dual-review human verification workflow. Each candidate is randomly assigned to at least two trained annotators, who inspect the clip, trajectory, draft expression, and candidate labels. The screening criteria are strict: reviewers must confirm that the expression explicitly describes the target through motion cues, including at least one motion-related term; that the clip exhibits observable UAV viewpoint change rather than a fixed-camera view; that the target can be unambiguously localized by a human while playing the video; and that the frame-level boxes match the referred target throughout the sequence. When necessary, reviewers are allowed to make limited edits to the draft expression, mainly to refine motion verbs, temporal connectors, and short motion descriptions or to remove minor ambiguity. Any edited sample is treated as a new version, and previous approvals are invalidated before independent re-verification. A sample is released only when the same final version is approved by at least two annotators; samples that remain ambiguous, cannot be reliably localized by human annotators, or fail the motion-centricity or trajectory-consistency checks are discarded. In addition to per-sample dual verification, we conduct a daily quality audit by randomly spot-checking 10% of each annotator's daily workload. A daily batch is accepted only when the audit pass rate exceeds 95%; otherwise, the corresponding annotations are returned for re-inspection. After verification, 22,225 clip-expression pairs are retained from 40,320 candidates, yielding an acceptance rate of 55.1%.

(a) Scene category **(b)** Target category

Figure 3: Word clouds of scene and target categories.



(a) Video length **(b)** Expression length

Figure 4: Distributions of video length and expression length in MoRe-UAV.

3.3 Dataset Analysis.

We conduct a statistical analysis of MoRe-UAV to examine its scale, diversity, and motion-aware properties. As shown in Table 1, existing UAV grounding datasets mainly emphasize appearance-centric or spatial-centric descriptions, even when video data are available. In contrast, MoRe-UAV explicitly focuses on motion-centric grounding in UAV videos. The final benchmark contains 22,225 video-expression pairs and 7,415,622 annotated frames. In addition

to frame-level annotations and referring expressions, each sample is annotated with a scene label and an object label, which we summarize into 28 scene categories and 30 target categories after verification. The scene categories cover diverse environments such as campus, park, mountain, river, and ocean, while the target categories are dominated by movable entities including pedestrians, vehicles, animals, and boats. Fig. 3 visualizes these verified label distributions and highlights the broad environmental coverage and object diversity of MoRe-UAV.

To complement the category-level statistics, we further quantify the motion-centric and dynamic viewpoint nature of MoRe-UAV using lightweight text-trajectory proxies. On the language side, every released expression contains at least one motion cue by construction, with an average of 4.1 motion-related terms per query; moreover, 39.4% of expressions involve multi-stage motion with temporal connectors such as "then" and "while", and 28.6% describe relative motion such as "follow", "pass", and "behind". On the trajectory side, the benchmark exhibits an average normalized trajectory length of 0.71, about 1.22 major direction changes per clip after trajectory smoothing, and a median bounding-box scale span of 4.38× across frames. These statistics indicate that MoRe-UAV does not reduce to near-static appearance matching, but instead requires reasoning about nontrivial target motion together with pronounced viewpoint-induced target variation.

Fig. 4 further summarizes the length statistics of the benchmark. MoRe-UAV covers clips of nontrivial temporal duration for UAV grounding rather than collapsing to near-static snippets, while the expression distribution is deliberately biased toward multi-word descriptions instead of very short keyword-style queries. On average, each referring expression contains approximately 11 words, and each clip lasts around 10 seconds with about 333 frames, which is sufficient to accumulate target motion and cross-view appearance changes over time.

To support standardized evaluation, we divide the finalized samples into training, validation, and test splits at the source video level, with 15,432, 2,571, and 4,222 samples respectively. All clips from the same source video are assigned to the same split to avoid cross-split leakage. We then construct the splits to be roughly proportional across the verified object categories while ensuring that every verified object label appears in all three splits. The resulting partition preserves the overall category distribution as much as possible while reducing split-induced category omission bias. Additional metadata summaries, annotation format examples, split statistics, and download/use instructions are provided on the project page.

4 Benchmark Baselines

To provide a stronger reference point for future work on MoRe-UAV, we introduce a lightweight benchmark baseline built upon a frozen Qwen2.5-VL backbone [?] with parameter-efficient tuning. The main challenge of MoRe-UAV lies in two aspects: motion-centric grounding cues in the query and drastic cross-view appearance changes caused by UAV ego-motion. To address them, we introduce two lightweight adapters: a **Motion-aware Prefix Adapter** (MPA), which emphasizes motion-sensitive semantics in the query, and a **Multi-view Alignment Adapter** (MVA), which captures temporally aligned target variations across consecutive frames. As shown in Fig. 5, MPA strengthens motion-aware language representation,

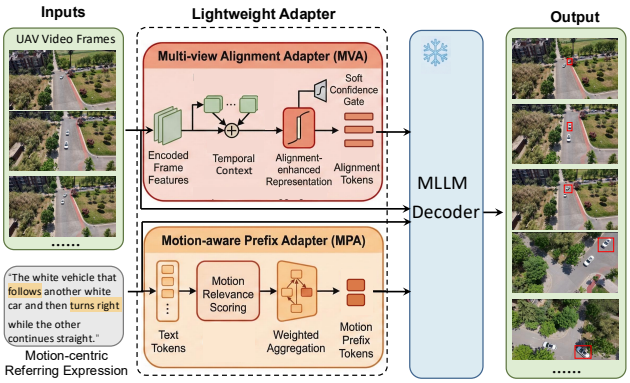


Figure 5: Overview of our baseline for motion-aware UAV video grounding. MPA emphasizes motion-sensitive query semantics, while MVA captures cross-view temporal alignment.

while MVA improves grounding robustness under dynamic UAV viewpoints. Further implementation details are available on our website and in the codebase.

4.1 Motion-aware Prefix Adapter (MPA)

Motion cues are crucial for distinguishing the referred target in UAV videos, yet directly relying on manually extracted motion words is often insufficient, since grounding usually depends on broader contextual composition. To address this, we design a Motion-aware Prefix Adapter (MPA) to emphasize motion-sensitive semantics in the query through weakly supervised token reweighting.

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 verbs and directional phrases [?] in the query. MPA uses a lightweight learnable scorer to produce motion-aware token weights $\alpha_i = \mathrm{softmax}(\mathbf{w}_m^{\top}\mathbf{l}_i)$, which are supervised by the pseudo-labels through

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

The motion-centric query representation is then obtained by weighted aggregation:

$$\mathbf{z}_m = \sum_{i=1}^{N} \alpha_i \mathbf{l}_i. \tag{2}$$

Finally, $\mathbf{z}_m$ is transformed into a small set of prefix prompts and prepended to the decoder input. In this way, pseudo-labels provide weak motion guidance, while the learnable weighting mechanism enables MPA to go beyond explicit motion word prompting and capture richer motion-sensitive semantics for grounding.

4.2 Multi-view Alignment Adapter (MVA)

To handle drastic viewpoint and appearance changes caused by UAV ego-motion, we introduce a Multi-view Alignment Adapter (MVA) that aligns target-related visual representations across consecutive frames. Given the input video sequence, MVA performs sequential alignment within the adapter, enabling temporally coherent feature aggregation while preserving the grounding focus of each frame under holistic video-level understanding.

Table 2: Comparison with baseline methods on MoRe-UAV.

Method	mIoU	Acc@0.5	Norm Prec.	AUC
STVG Methods				
Co-grounding Model [?]	5.84	5.10	6.20	5.55
CG-STVG [?]	7.95	7.10	9.82	8.33
MLLM + Tracking				
MiniCPM-V + Tracker [?]	8.23	6.58	7.95	8.61
Qwen2.5-VL + Tracker [?]	9.48	10.02	11.25	9.90
MLLM				
MiniCPM-V [?]	10.18	10.93	11.84	10.59
Qwen2.5-VL [?]	12.00	10.95	14.52	12.36
Ours				
MoRe-UAV (MiniCPM)	19.12	18.91	21.34	19.45
MoRe-UAV (Qwen)	**21.22**	**20.25**	**25.60**	**21.45**
Human performance	89.33	91.13	90.51	89.41

Let $\mathbf{v}_t$ denote the visual representation of frame t, and let $\mathbf{s}_{t-1}$ denote the propagated alignment state from adjacent temporal context. MVA combines the current frame representation with the propagated context to produce an aligned representation

$$\mathbf{c}_t = f_a([\mathbf{v}_t; \mathbf{s}_{t-1}]), \tag{3}$$

where $f_a(\cdot)$ denotes a lightweight attention-guided fusion and alignment function. Since propagated context may become unreliable under blur, occlusion, or abrupt camera motion, we further introduce a soft confidence gate and compute the enhanced representation as

$$\mathbf{h}_t = g_t \cdot \mathbf{c}_t + (1 - g_t) \cdot \mathbf{v}_t, \quad g_t = \sigma\big(f_g([\mathbf{v}_t; \mathbf{c}_t])\big), \tag{4}$$

where $f_g(\cdot)$ is a lightweight gating function. The gate adaptively controls how much aligned temporal context should contribute to the current frame. The resulting representations are then transformed into alignment prompts and fed into the decoder, improving grounding consistency under dynamic UAV viewpoints.

5 Experiments

We benchmark three categories of methods on MoRe-UAV: spatio-temporal video grounding methods, direct MLLM baselines, and MLLM+tracking pipelines.

5.1 Implementation Details

We report mIoU, Acc@0.5, Norm Precision, and IoU AUC following prior video grounding and tracking benchmarks [? ? ?]. All methods are trained or adapted on the same MoRe-UAV split. For direct MLLM baselines, we use the same frozen backbones and parameter-efficient tuning budget as our method, but without MPA and MVA. For STVG methods, including Co-grounding [?] and CG-STVG [?], we follow the official training protocols when applicable and evaluate them on the same split. Frames without visible targets are annotated as empty and evaluated over the full video sequence. We additionally report human performance on a randomly sampled 20% subset of the test set. Human annotators were allowed to pause and replay videos during evaluation, and their annotations were evaluated using the same protocol and metrics as model predictions. Detailed prompts, preprocessing rules, optimization and inference settings, evaluation protocol, and command-line examples are provided in the supplementary material and code repository.

Table 3: Ablation study of MPA and MVA on the Qwen2.5-VL backbone.

Method	mIoU	Acc@0.5	Norm Prec.	AUC
Qwen2.5-VL (baseline)	12.00	10.95	14.52	12.36
MPA (Motion Modeling)				
+ Motion Word Prompt (NLTK verbs)	14.35	13.80	17.92	14.70
+ MPA (ours)	**17.86**	**16.95**	**21.03**	**18.10**
MVA (Multi-view Alignment)				
+ Iterative box guidance	15.02	14.20	18.11	15.38
+ MVA (ours)	**18.74**	**17.82**	**22.65**	**19.02**
Combined				
+ MPA + MVA (ours)	21.22	20.25	25.60	21.45

5.2 Results and Analysis

As shown in Table 2, all existing baselines perform poorly on MoRe-UAV, highlighting the difficulty of motion-aware visual grounding in UAV videos. For MLLM+tracking pipelines, the MLLM is used only for initial target identification, while subsequent frames rely on a tracker, making them particularly vulnerable to drift under target motion and drastic viewpoint changes. Direct MLLM grounding also remains far below human performance. The large gap between current models and human performance further indicates that the benchmark remains far from saturated. Although draft expressions are initialized by a local Qwen2.5-VL assistant, all released annotations are manually reviewed and revised when necessary. Our method consistently improves both Qwen and MiniCPM backbones, suggesting that motion-sensitive query modeling and cross-view temporal alignment are both important for this benchmark and that the gains are not limited to a single model family. Additional qualitative examples and failure cases are provided in the supplementary material.

We also conduct ablations on the Qwen2.5-VL backbone to evaluate the effectiveness of MPA and MVA. As shown in Table 3, a simple Motion Word Prompt already improves the baseline, confirming the importance of motion cues, but its gains remain limited because it does not model broader contextual dependencies. In contrast, MPA yields substantially larger improvements through learnable motion-aware token aggregation. We also compare MVA with a naive temporal strategy that appends the previous prediction as a textual hint for the next frame. While this strategy brings moderate gains, MVA performs better by explicitly modeling cross-view temporal consistency. Combining MPA and MVA gives the best results, showing that the two components are complementary.

6 Conclusion

We introduced **MoRe-UAV**, a large-scale benchmark for motion-aware visual grounding in UAV videos, featuring motion-centric queries, moving targets, and strong viewpoint changes caused by UAV ego-motion. We established benchmark evaluations across STVG methods, MLLMs, and MLLM+tracking pipelines, and provided a stronger reference baseline with MPA and MVA. The substantial gap between current models and human performance shows that MoRe-UAV remains a challenging testbed for future research on motion-aware and multi-view grounding in UAV videos.

References

[] Amnesty International and Access Now. 2018. The Toronto Declaration: Protecting the Right to Equality and Non-Discrimination in Machine Learning Systems. https://www.torontodeclaration.org/declaration-text/english/ Accessed: 2026-03-30.

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

[] Steven Bird. 2006. NLTK: the natural language toolkit. In *Proceedings of the COLING/ACL 2006 interactive presentation sessions*. 69–72.

[] Sining Cheng, Jiaxian Qin, Yuanyuan Chen, and Mingzhu Li. 2022. Moving target detection technology based on UAV Vision. *Wireless Communications and Mobile Computing* 2022, 1 (2022), 5443237.

[] Jiajun Deng, Zhengyuan Yang, Tianlang Chen, Wengang Zhou, and Houqiang Li. 2021. Transvg: End-to-end visual grounding with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*. 1769–1779.

[] Henghui Ding, Chang Liu, Shuting He, Xudong Jiang, and Chen Change Loy. 2023. Mevis: A large-scale benchmark for video segmentation with motion expressions. In *Proceedings of the IEEE/CVF international conference on computer vision*. 2694–2703.

[] Xin Gu, Heng Fan, Yan Huang, Tiejian Luo, and Libo Zhang. 2024. Context-guided spatio-temporal video grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 18330–18339.

[] Wenyi Hong, Yean Cheng, Zhuoyi Yang, Weihan Wang, Lefan Wang, Xiaotao Gu, Shiyu Huang, Yuxiao Dong, and Jie Tang. 2025. Motionbench: Benchmarking and improving fine-grained video motion understanding for vision language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 8450–8460.

[] Deyi Ji, Lanyun Zhu, Siqi Gao, Qi Zhu, Yiru Zhao, Peng Xu, Yue Ding, Hongtao Lu, Jieping Ye, Feng Wu, et al. 2026. View-centric multi-object tracking with homographic matching in moving uav. *IEEE Transactions on Geoscience and Remote Sensing* (2026).

[] Yang Jin, Zehuan Yuan, Yadong Mu, et al. 2022. Embracing consistency: A one-stage approach for spatio-temporal video grounding. *Advances in Neural Information Processing Systems* 35 (2022), 29192–29204.

[] Lingdong Kong, Dongyue Lu, Ao Liang, Rong Li, Yuhao Dong, Tianshuai Hu, Lai Xing Ng, Wei Tsang Ooi, and Benoit R Cottereau. 2025. Talk2Event: Grounded understanding of dynamic scenes from event cameras. *arXiv preprint arXiv:2507.17664* (2025).

[] N Murali Krishna, Ramidi Yashwanth Reddy, Mallu Sai Chandra Reddy, Kasibhatla Phani Madhav, and Gaikwad Sudham. 2021. Object detection and tracking using YOLO. In *2021 Third International Conference on Inventive Research in Computing Applications (ICIRCA)*. IEEE, 1–7.

[] Hengyou Li, Xinyan Liu, and Guorong Li. 2024. A benchmark for UAV-view natural language-guided tracking. *Electronics* 13, 9 (2024), 1706.

[] Yan Li, Weiwei Guo, Xue Yang, Ning Liao, Dunyun He, Jiaqi Zhou, and Wenxian Yu. 2024. Toward Open Vocabulary Aerial Object Detection with CLIP-Activated Student-Teacher Learning. arXiv:2311.11646 [cs.CV] https://arxiv.org/abs/2311.11646

[] Yunhao Li, Xiaoqiong Liu, Luke Liu, Heng Fan, and Libo Zhang. 2025. Lamot: Language-guided multi-object tracking. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 6816–6822.

[] Junli Liu, Qizhi Chen, Zhigang Wang, Yiwen Tang, Yiting Zhang, Chi Yan, Dong Wang, Xuelong Li, and Bin Zhao. 2025. Aerialvg: A challenging benchmark for aerial visual grounding by exploring positional relations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 5177–5187.

[] Shuai Liu, Xin Li, Huchuan Lu, and You He. 2022. Multi-object tracking meets moving UAV. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8876–8885.

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

[] Matthias Mueller, Neil Smith, and Bernard Ghanem. 2016. *UAV123 Dataset*. https://cemse.kaust.edu.sa/ivul/uav123

[] Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2025. Qwen2.5 Technical Report. arXiv:2412.15115 [cs.CL] https://arxiv.org/abs/2412.15115

[] Arka Sadhu, Kan Chen, and Ram Nevatia. 2020. Video object grounding using semantic roles in language description. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10417–10427.

[] Guorui Song, Guocun Wang, Zhe Huang, Jing Lin, Xuefei Zhe, Jian Li, and Haoqian Wang. 2025. Towards Fine-Grained Human Motion Video Captioning. In *Proceedings of the 33rd ACM International Conference on Multimedia*. 846–855.

[] Jianbo Song, Hong Zhang, Yachun Feng, Hanyang Liu, and Yifan Yang. 2025. Language-guided Visual Tracking: Comprehensive and Effective Multimodal Information Fusion. *ACM Transactions on Multimedia Computing, Communications and Applications* 21, 10 (2025), 1–23.

[] Sijie Song, Xudong Lin, Jiaying Liu, Zongming Guo, and Shih-Fu Chang. 2021. Co-grounding networks with semantic attention for referring expression comprehension in videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1346–1355.

[] Zhichao Sun, Yepeng Liu, Zhiling Su, Huachao Zhu, Yuliang Gu, Yuda Zou, Zelong Liu, Gui-Song Xia, Bo Du, and Yongchao Xu. 2025. RefDrone: A Challenging Benchmark for Referring Expression Comprehension in Drone Scenes. arXiv:2502.00392 [cs.CV] https://arxiv.org/abs/2502.00392

[] Zongheng Tang, Yue Liao, Si Liu, Guanbin Li, Xiaojie Jin, Hongxu Jiang, Qian Yu, and Dong Xu. 2021. Human-centric spatio-temporal video grounding with visual transformers. *IEEE Transactions on Circuits and Systems for Video Technology* 32, 12 (2021), 8238–8249.

[] Jingchao Wang, Kaiwen Zhou, Zhijian Wu, Kunhua Ji, Dingjiang Huang, and Yefeng Zheng. 2025. VPTracker: Global Vision-Language Tracking via Visual Prompt and MLLM. *arXiv preprint arXiv:2512.22799* (2025).

[] Mengmeng Wang, Jiazheng Xing, Jianbiao Mei, Yong Liu, and Yunliang Jiang. 2023. Actionclip: Adapting language-image pretrained models for video action recognition. *IEEE transactions on neural networks and learning systems* 36, 1 (2023), 625–637.

[] Wei Wang, Junyu Gao, and Changsheng Xu. 2021. Weakly-supervised video object grounding via stable context learning. In *Proceedings of the 29th ACM International Conference on Multimedia*. 760–768.

[] Wei Wang, Junyu Gao, and Changsheng Xu. 2022. Weakly-supervised video object grounding via learning uni-modal associations. *IEEE Transactions on Multimedia* 25 (2022), 6329–6340.

[] Syed Talal Wasim, Muzammal Naseer, Salman Khan, Ming-Hsuan Yang, and Fahad Shahbaz Khan. 2024. Videogrounding-dino: Towards open-vocabulary spatio-temporal video grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 18909–18918.

[] Guoting Wei, Xia Yuan, Yang Zhou, Haizhao Jing, Yu Liu, Xianbiao Qi, Chunxia Zhao, Haokui Zhang, and Rong Xiao. 2026. Open-Text Aerial Detection: A Unified Framework For Aerial Visual Grounding And Detection. arXiv:2602.07827 [cs.CV] https://arxiv.org/abs/2602.07827

[] Xiyang Wu, Zongxia Li, Jihui Jin, Guangyao Shi, Gouthaman KV, Vishnu Raj, Nilotpal Sinha, Jingxi Chen, Fan Du, and Dinesh Manocha. 2025. MASS: Motion-Aware Spatial-Temporal Grounding for Physics Reasoning and Comprehension in Vision-Language Models. *arXiv preprint arXiv:2511.18373* (2025).

[] Jianqiang Xiao, Yuexuan Sun, Yixin Shao, Boxi Gan, Rongqiang Liu, Yanjin Wu, Weili Guan, and Xiang Deng. 2025. Uav-on: A benchmark for open-world object goal navigation with aerial agents. In *Proceedings of the 33rd ACM International Conference on Multimedia*. 13023–13029.

[] Linhui Xiao, Xiaoshan Yang, Xiangyuan Lan, Yaowei Wang, and Changsheng Xu. 2025. Towards visual grounding: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2025).

[] Chaocan Xue, Bineng Zhong, Qihua Liang, Yaozong Zheng, Ning Li, Yuanliang Xue, and Shuxiang Song. 2025. Similarity-Guided Layer-Adaptive Vision Transformer for UAV Tracking. arXiv:2503.06625 [cs.CV] https://arxiv.org/abs/2503.06625

[] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. 2022. Tubedetr: Spatio-temporal video grounding with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 16442–16453.

[] Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. 2024. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800* (2024).

[] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. 2016. Modeling context in referring expressions. In *European conference on computer vision*. Springer, 69–85.

[] Yang Zhan and Yuan Yuan. 2025. Where Does It Exist from the Low-Altitude: Spatial Aerial Video Grounding. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

[] Yang Zhan and Yuan Yuan. 2026. UAVBench and UAVIT-1M: Benchmarking and Enhancing MLLMs for Low-Altitude UAV Vision-Language Understanding. arXiv:2603.14336 [cs.CV] https://arxiv.org/abs/2603.14336

[] Chunhui Zhang, Guanjie Huang, Li Liu, Shan Huang, Yinan Yang, Xiang Wan, Shiming Ge, and Dacheng Tao. 2022. WebUAV-3M: A benchmark for unveiling the power of million-scale deep UAV tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, 7 (2022), 9186–9205.

[] Zhu Zhang, Zhou Zhao, Yang Zhao, Qi Wang, Huasheng Liu, and Lianli Gao. 2020. Where does it exist: Spatio-temporal video grounding for multi-form sentences. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10668–10677.

[] Yue Zhou, Jue Chen, Zilun Zhang, Penghui Huang, Ran Ding, Zhentao Zou, PengFei Gao, Yuchen Wei, Ke Li, Xue Yang, et al. 2026. DVGBench: Implicit-to-explicit visual grounding benchmark in UAV imagery with large vision–language models. *ISPRS Journal of Photogrammetry and Remote Sensing* 232 (2026), 831–847.

[] Yue Zhou, Mengcheng Lan, Xiang Li, Litong Feng, Yiping Ke, Xue Jiang, Qingyun Li, Xue Yang, and Wayne Zhang. 2025. GeoGround: A Unified Large Vision-Language Model for Remote Sensing Visual Grounding. arXiv:2411.11904 [cs.CV] https://arxiv.org/abs/2411.11904

[] Pengfei Zhu, Longyin Wen, Dawei Du, Xiao Bian, Heng Fan, Qinghua Hu, and Haibin Ling. 2021. Detection and tracking meet drones challenge. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44, 11 (2021), 7380–7399.