CAN VISION-LANGUAGE MODELS CONTROL DRONES? DIAGNOSING VIEWPOINT-AWARE SPATIAL GROUNDING

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision-language models (VLMs) are increasingly viewed as promising interfaces for aerial systems, where they could translate human goals and visual observations into drone commands. However, it remains unclear whether they can reliably map image-grounded spatial relations to directions defined in the aircraft movement frame. To address this gap, we introduce DroneBench, a diagnostic benchmark for one-step command prediction from an aerial image and a result-oriented task goal. DroneBench contains 1,350 paired images and tasks across four operational scenarios and two viewpoints, with each command decomposed into speed, aircraft-centered direction, and action for component-wise evaluation. Experiments with 15 VLMs show that models generally recognize task-relevant targets and choose plausible operational actions, yet remain substantially less reliable at direction grounding, especially for top-down views. We further evaluate five models on 50 matched forward-facing/top-down pairs collected during physical drone flights and observe the same failure pattern on onboard imagery. These results identify viewpoint-aware coordinate conversion as a central obstacle to reliable aerial command selection and establish DroneBench as a practical benchmark for advancing spatial reasoning in aerial VLMs.

1 INTRODUCTION

The low-altitude economy is emerging as an important growth area for logistics, inspection, emergency response, urban management, and environmental monitoring (State Council of the People's Republic of China, 2024). Unmanned aerial vehicles (UAVs) provide the mobile sensing platform for many of these services, while vision-language models (VLMs) offer a promising interface for turning human goals and visual observations into task decisions. Their combination could make aerial systems easier to instruct and more adaptable across missions, provided that a model can reliably convert what it sees into commands meaningful to the aircraft. However, aerial VLM development lacks a diagnostic platform for testing the conversion from visual evidence to aircraft-centered movement. A drone command is expressed in the aircraft movement frame, whereas the target relation is observed in the image frame. Forward-facing views provide perspective and heading cues, while top-down views suppress many of them and require a different interpretation of image position. Existing aggregate evaluations cannot reveal whether a model failed to recognize the task-relevant target or recognized it but selected the wrong movement axis. This missing diagnosis limits our ability to measure and improve control-relevant spatial grounding.

We introduce **DroneBench**, a diagnostic benchmark that maps an aerial image and a result-oriented task goal to a one-step command consisting of speed, aircraft-centered direction, and action. The model stands where a human operator stands. It receives only the camera image and the task goal, and must itself determine the decision-layer components of speed, direction, and action. Its 1,350 examples cover four operational scenarios and two viewpoints, and its component-wise scoring separates recognition and operation selection from movement-frame grounding. Across 15 VLMs, direction is the weakest component and degrades most under top-down views. A matched real-flight extension with 100 onboard images reproduces the same component-specific failure pattern.

Our contributions are threefold:

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.

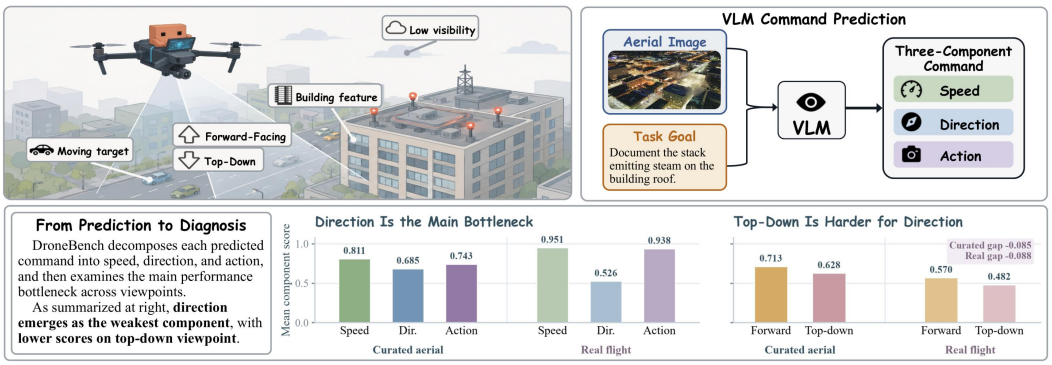


Figure 1: DroneBench evaluates viewpoint-aware spatial grounding for aerial command prediction. Given an aerial image and a result-oriented goal, a VLM predicts speed, direction, and action. The curated benchmark and matched real-flight extension jointly test whether component behavior persists across viewpoints and image sources.

- We build DroneBench, a 1,350-example diagnostic platform for one-step aerial command prediction across four operational scenarios, with viewpoint-conditioned construction and component-wise evaluation of speed, direction, and action.
- We evaluate 15 mainstream VLMs and show that current models remain unreliable for this task: direction grounding is the primary bottleneck for most models, especially under top-down views, even when speed and action are often correct.
- We validate the diagnosis on 50 matched forward-facing/top-down pairs that each share one task, collected during physical drone flights, where the same separation between strong operation selection and weak direction grounding persists.

2 Related Work

Multimodal and embodied evaluation. DroneBench joins multimodal evaluation with embodied decision making. MMBench, SEED-Bench, and MM-Vet measure broad visual perception and reasoning (Liu et al., 2023b; Li et al., 2023; Yu et al., 2024), GeoChat and LHRS-Bot ground language in remote-sensing imagery (Kuckreja et al., 2023; Muhtar et al., 2024), PaLM-E and RT-2 tie vision-language reasoning to robotic action (Driess et al., 2023; Brohan et al., 2023), and DriveBench probes decision reliability in driving scenes (Xie et al., 2025). Such evaluations, however, stop before aircraft motion, and ground vehicles do not expose the free-flight axes, vertical movement, and viewpoint changes that mark UAV operation.

Aerial perception and simulation. UAVDT and VisDrone benchmark detection and tracking under altitude, viewpoint, camera-motion, and scale variation (Du et al., 2018; Zhu et al., 2022), and AirSim and Flightmare provide simulation environments for visual autonomy and agile quadrotor learning (Shah et al., 2018; Song et al., 2020). AeroVerse evaluates embodied capabilities, from scene awareness and spatial reasoning to navigation, planning, and motion decisions, as task performance inside a simulated interaction loop (Yao et al., 2024), where a failure can stem from perception or planning as readily as from the conversion of one observed relation into an aircraft-frame command, and cannot be attributed to a specific command component.

Language-guided aerial intelligence. CityNav and OpenUAV study language-goal navigation and in-flight assistance (Lee et al., 2024; Chen et al., 2024), while CityAVOS, AirCopBench, and AeroDuo address urban object search and collaborative navigation (Chen et al., 2025; Liu et al., 2025; Wu et al., 2025). AerialVLA, WorldVLN, and ImagineUAV develop learned policies or world-action models for closed-loop UAV navigation (Huang et al., 2026; Zhao et al., 2026; Liu et al., 2026), and ActiveFly-Bench evaluates active visual understanding through flight-based interaction (Deng et al., 2026). These systems score end-to-end behavior, in which perception, planning, and control are entangled and a downstream planner can compensate for a weak spatial decision. DroneBench instead reproduces the decision situation of a human operator acting from the camera feed alone. The model receives the aerial image, the result-oriented task, and the fixed command specification,

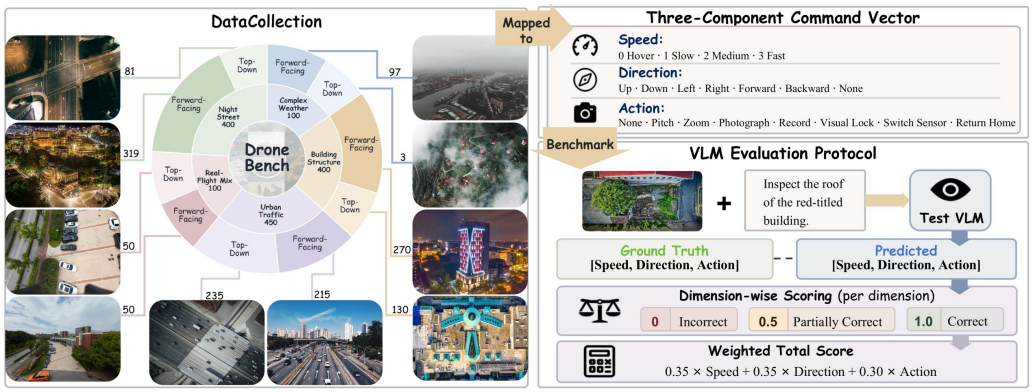


Figure 2: Construction and evaluation pipeline of DroneBench. The main set contains 1,350 examples stratified by scenario and viewpoint. A separate real-flight extension adds 50 manually labeled forward-facing/top-down pairs (100 images), with one shared task per pair and image-specific reference commands. Predictions are scored independently as incorrect (0), partially correct (0.5), or correct (1), then combined into a weighted total.

with no telemetry, pose estimate, or interactive feedback, and must itself resolve, in one step, the three decision-layer components of speed, direction, and action. Failure is therefore attributable to a specific component, and matched forward-facing/top-down pairs, in curated and real-flight imagery, further isolate the viewpoint variable that no trajectory- or interaction-level metric exposes.

Fine-grained command evaluation. The evaluation metric decides which failures are visible. Exact match treats a plausible near miss and an unrelated command identically, hiding whether the model has partially grasped the task. Following reference-guided LLM evaluation (Liu et al., 2023a), DroneBench scores speed, direction, and action using explicit criteria and scene rationales. This decomposition separates errors in operation selection from errors in movement-frame grounding.

3 DroneBench

3.1 Task Definition

Figure 2 summarizes the pipeline. Each DroneBench instance is a tuple (I, T, G) with aerial image I, task goal T, and ground truth G. The model receives I and T and predicts

$$\hat{G} = [\hat{s}, \hat{d}, \hat{a}], \tag{1}$$

where $\hat{s}$ is speed, $\hat{d}$ is movement direction, $\hat{a}$ is an action. The goal specifies a result, such as inspecting a roof. Each predicted command contains one speed level, one direction, and one action.

Speed. Speed has four levels: 0 (hover), 1 (slow), 2 (medium), and 3 (fast), with construction ranges of $[0, 0.5)$, $[0.5, 6)$, $[6, 15)$, and $[15, 21]$ meters per second. It reflects stability and urgency: inspection and photography favor slower motion to limit blur, while fast is reserved for urgency.

Direction. Direction is selected from {Up, Down, Left, Right, Forward, Backward, None}; "None" is valid only at zero speed. Direction is defined in an aircraft-centered movement frame.

Action. The action set is {None, Pitch, Visual Lock, Zoom, Photograph, Record, Return Home, Switch Sensor}. It covers camera and flight goals without introducing low-level controls.

3.2 Viewpoint-Dependent Direction Rules

Because camera orientation changes the maneuver, construction uses viewpoint-specific rules.

Forward-facing views. These images contain oblique geometry: the horizon or sky may be visible, roads converge, and facades provide depth cues. Targets near the optical axis are assigned Forward. Large horizontal displacement yields Left or Right, while predominantly vertical displacement yields Up or Down. Backward is used when the task requires retreating.

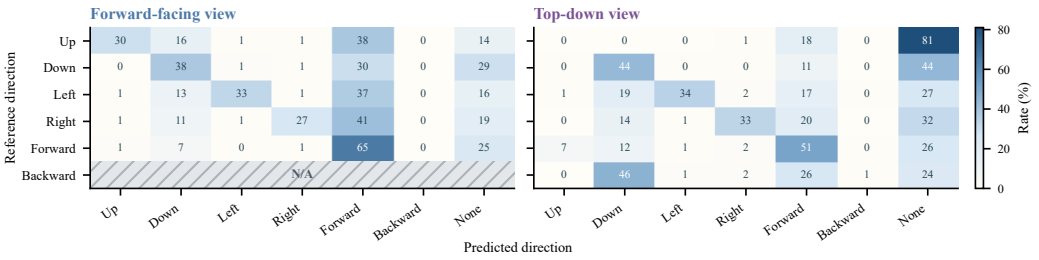


Figure 3: Viewpoint-conditioned direction confusion over 15 VLMs. Rows denote references, columns denote canonical predictions, entries are row-normalized. Top-down views produce more substitutions among *Down*, *Backward*, *Forward*, *None*, exposing uncertainty in mapping image-plane relations to aircraft-centered motion. The forward-facing subset contains no *Backward*.

Top-down views. For a nadir camera, the image center is the ground point below the drone, and image top defines forward heading. A target above, below, left, or right of center maps to Forward, Backward, Left, or Right. A central target may require Down for inspection; Up is used for wider coverage. The rule blocks are asymmetric. A forward-facing image does not reveal the region behind the aircraft, so Backward is used only when the goal requires increasing separation. A top-down image exposes both sides of the forward axis, allowing the lower half to represent the region behind the aircraft, while Up and Down combine coverage or inspection requirements with target location. These engineering rules produce an actionable command from a static image.

3.3 Image Collection and Scenario Coverage

DroneBench contains 1,350 aerial images in four scenarios: Building Structure (400; rooftops and facades), Urban Traffic (450; roads, vehicles and flows), Night Street (400 low-light scenes), and Complex Weather (100 visibility-degraded scenes). Categories guide construction and analysis. The set has 901 forward-facing and 449 top-down examples; Figure 2 shows both distributions.

3.4 Task and Reference Construction

We use a two-stage, view-conditioned pipeline. Stage 1 performs coarse viewpoint classification. In Stage 2, the predicted viewpoint first selects the corresponding direction-rule block; an annotator, aided by a vision model, then produces a goal, category, command, and rationale using that block. This separation avoids mixing incompatible mappings and preserves the viewpoint label for stratified analysis. Tasks state what to document, track, or inspect, never how to move: direction and speed terms such as “left,” “down,” “slowly,” and “quickly” are prohibited. Each task depends on visible evidence, requires one command, and is validated against a fixed JSON schema. Each example in the main set and the real-flight extension carries exactly one reference command, mirroring a human operator who faces a single returned frame and commits to one value for speed, one for direction, and one for action. Annotators record the single best command; a set of acceptable alternatives is not part of the protocol. Near-miss alternatives receive partial credit, and the judge grants 0.5 for a useful non-exact speed, an adjacent or non-opposite movement direction, or a related operation.

3.5 Matched Real-Flight Extension

To test the pattern on imagery captured during physical flights, we collect 100 onboard images in 50 matched pairs using a DJI Avata 2, with the aircraft and camera system documented in its user manual (DJI, 2024). Each pair depicts the same target or local scene through one forward-facing and one top-down image. A human annotator assigns the view labels and writes one result-oriented task shared by both images. Because both images serve the same task, the action label is expected to stay the same across the pair, while direction (and occasionally speed) follows the viewpoint rule of each image and may differ between the two views. All 50 pairs share the action label, 27 require different directions, and three require different speeds. After fixing the task, viewpoint, and command, a vision-capable annotation model completes each rationale, which is reviewed before export. The extension reuses the main action space, task principles, direction rules, and rationale format.

4 Evaluation Protocol

4.1 Test-Model Prompting

Models receive the same image, task goal, command definitions, and JSON format (Figure 2). No telemetry, camera pose, or interactive feedback is provided. The decision rests on the visual channel and the task, as it would for a human operator. Beyond speed ranges and the constraint that Direction = None applies only at zero speed, the prompt gives no viewpoint-specific rule. Removing the rules is deliberate; the model must infer the image-to-aircraft transformation from visual evidence alone. It reads perspective and horizon cues in forward-facing views and image-position cues in top-down views under the stated heading convention, just as a human operator does from a live camera feed. A parser preserves each predicted component for diagnosis.

4.2 Component-Wise Scoring

An independent judge compares the prediction with the task, reference command, and construction rationale. The rationale records the target or scene condition, spatial relation, and why the reference satisfies the task. Each component receives a score in $\{0, 0.5, 1\}$. An exact match receives 1. A different speed may receive 0.5 if it preserves stability and urgency. Direction receives 0.5 when a non-opposite heading retains movement toward the target; movement away receives 0. An alternative action receives 0.5 only for a related goal, such as zooming rather than photographing to read a distant sign. The judge grants exact matches immediately and considers semantic equivalence only when labels differ. The total score is

$$S = \alpha S_{\text{speed}} + \beta S_{\text{direction}} + \gamma S_{\text{action}}. \tag{2}$$

We set $\alpha = 0.35$, $\beta = 0.35$, and $\gamma = 0.30$, giving speed and direction equal weight and action slightly less. This reflects an asymmetry in how each error affects task completion: an error in either speed or direction alone can make the task fail outright, since both are needed to reach and stay with the target. For example, when tracking a moving vehicle, a wrong speed lets the target escape the frame and a wrong direction points the drone away from it; either error alone is enough to break tracking. An action error is more forgiving: recording in place of photographing, for instance, still yields usable visual evidence of the target, so it often preserves partial task utility even when not exactly correct. We report each component, the three-component weighted total, and full-command exact match, which requires three scores of 1. Partial credit captures useful non-exact alternatives, while full-command exact match preserves a strict measure of joint command correctness.

5 Experimental Results

5.1 Models and Implementation

Evaluated models. We evaluate 15 VLMs: Doubao-Seed-1.6-Vision (ByteDance Seed Team, 2025), Gemini 2.5 Flash (Gemini Team, 2025), GLM-4.6V (Z.ai, 2025), GLM-5V-Turbo (GLM Team, 2026), Gemma 4 31B (Gemma Team, 2026), GPT-5.4 and GPT-5.4 mini (OpenAI, 2026a;b), Grok 4.3 (xAI, 2026), Kimi K2.6 (Moonshot AI, 2026), Llama 3.2 11B Vision (Meta AI, 2024), MiMo-V2.5 (Xiaomi MiMo Team, 2026), Mistral Medium 3 (Mistral AI, 2025), Qwen3-27B (Qwen Team, 2025), Qwen3.5-397B-A17B and Qwen3.5-9B (Qwen Team, 2026). Each receives 1,350 examples. The real-flight test evaluates five stronger models: GLM-5V-Turbo, Gemma 4 31B, Kimi K2.6, Mistral Medium 3, and Qwen3.5-397B-A17B on 100 real-flight images.

Implementation and compute. Evaluations ran on Windows 11 with Python 3.11, using the OpenAI-compatible client (`openai`$\geq$1.0.0) and Pillow $\geq$10.0.0. All VLMs and the fixed GPT-4o judge (OpenAI, 2024) are provider-hosted; no training or fine-tuning is performed.

Reporting. We report model, component, scenario, and viewpoint averages, together with a sampled human reference. Per-viewpoint and per-scenario breakdowns are reported in Tables 2 and 3.

5.2 Benchmark Results

Across 20,250 evaluations, the mean weighted score is 0.7465. Table 1 reports each model and the human reference. GLM-5V-Turbo ranks first at 0.8364, followed by Mistral Medium 3 at 0.8117,

Table 1: Overall results on DroneBench. Speed, Dir., and Action are mean component scores; Total is their weighted combination. The Exact match block gives the share of predictions with that component exactly correct; Full is the full-command exact-match rate (%). Models are evaluated on all 1,350 examples; Human uses 213 samples. Boldface marks the best model result.

Model	Mean score				Exact match (%)			
	Speed↑	Dir.↑	Action↑	Total↑	Speed↑	Dir.↑	Action↑	Full↑
Human	0.939	0.927	0.923	0.930	87.8%	87.8%	87.3%	71.4%
GLM-5V-Turbo	**0.910**	**0.765**	0.834	**0.836**	**82.1%**	**57.9%**	**74.9%**	**38.0%**
Mistral-Medium-3	0.904	0.686	**0.851**	0.812	80.9%	45.6%	74.4%	27.3%
Qwen3.5-397B-A17B	0.855	0.752	0.800	0.803	71.2%	55.6%	70.6%	34.2%
Gemma-4-31B	0.843	0.741	0.821	0.801	68.7%	51.6%	74.1%	33.9%
Kimi-K2.6	0.831	0.706	0.796	0.776	66.4%	46.7%	72.4%	29.1%
Qwen3.5-9B	0.831	0.740	0.743	0.773	67.0%	53.8%	67.8%	30.2%
Doubao-Seed-1.6-Vision	0.873	0.715	0.695	0.765	75.4%	51.4%	64.1%	25.5%
Gemini-2.5-Flash	0.813	0.685	0.765	0.754	63.9%	43.7%	69.9%	24.1%
Qwen3-27B	0.787	0.676	0.759	0.740	58.6%	40.1%	68.8%	22.6%
GPT-5.4	0.739	0.677	0.806	0.738	49.3%	38.8%	71.9%	24.4%
GLM-4.6V	0.774	0.666	0.746	0.728	55.7%	37.8%	67.3%	21.3%
Grok-4.3	0.733	0.661	0.744	0.711	47.6%	36.2%	67.9%	21.2%
MiMo-V2.5	0.752	0.640	0.657	0.684	54.3%	33.7%	60.2%	20.1%
Llama-3.2-11B-Vision	0.806	0.614	0.578	0.671	65.5%	37.6%	50.7%	13.0%
GPT-5.4-mini	0.715	0.550	0.549	0.607	46.4%	20.5%	48.8%	11.0%

and Qwen3.5-397B-A17B at 0.8026; the lowest score is 0.6071. This 0.2293 range separates current VLMs on this task. The contextual human reference reaches 0.9299 on 213 sampled tasks.

Direction as the main bottleneck. The component averages are 0.8111 for speed, 0.6850 for direction, and 0.7430 for action. Direction is exactly correct in 43.41% of evaluations, compared with 63.54% for speed and 66.93% for action, and is the weakest component for 12 of 15 models. Partial credit is common for direction: 50.18% of predictions receive 0.5, while only 6.41% receive 0. Models therefore tend to choose plausible but geometrically imprecise motion. Figure 3 shows the recurring substitutions: Forward is often replaced by None, Right by Forward, and Down by None or Forward. These errors indicate that models often locate the relevant region but do not resolve whether its image-plane displacement implies lateral motion or altitude change.

Average totals range from 0.6827 in Complex Weather to 0.7608 in Building Structure. Complex Weather combines the highest direction score (0.7333) with the lowest action score (0.4163), whereas Urban Traffic has the highest action score (0.8265) and lowest speed score (0.7416). Ta-

Table 2: Results stratified by operational scenario. Model subsets contain 400 Building Structure, 100 Complex Weather, 400 Night Street, and 450 Urban Traffic examples; Human subsets contain 60, 25, 60, and 68 examples. Overall is the weighted score over the complete evaluation set.

Model	Building Structure			Complex Weather			Night Street			Urban Traffic			Overall
	Speed↑	Dir.↑	Action↑	Speed↑	Dir.↑	Action↑	Speed↑	Dir.↑	Action↑	Speed↑	Dir.↑	Action↑	Total↑
Human	1.000	0.933	0.908	0.960	0.960	0.760	0.942	0.900	0.975	0.875	0.934	0.949	0.930
GLM-5V	0.985	0.759	0.800	**0.955**	0.805	0.440	**0.917**	**0.770**	0.876	0.826	**0.757**	0.916	**0.836**
Mistral-M3	**0.989**	0.657	0.757	0.945	0.785	**0.565**	0.887	0.693	**0.899**	**0.833**	0.682	**0.957**	0.812
Qwen3.5-397B	0.961	**0.789**	0.766	0.920	0.790	0.515	0.833	0.735	0.829	0.766	0.727	0.869	0.803
Gemma-4-31B	0.917	0.733	**0.831**	0.940	0.795	0.490	0.809	0.741	0.830	0.786	0.738	0.877	0.801
Kimi-K2.6	0.945	0.700	0.719	0.915	0.775	0.515	0.825	0.701	0.848	0.716	0.699	0.881	0.776
Qwen3.5-9B	0.965	0.745	0.748	0.905	0.805	0.355	0.762	0.757	0.759	0.757	0.707	0.811	0.773
Doubao-1.6V	0.983	0.686	0.520	0.950	**0.835**	0.360	0.829	0.725	0.800	0.799	0.706	0.832	0.765
Gemini-2.5F	0.920	0.700	0.611	0.860	0.690	0.485	0.776	0.690	0.856	0.741	0.667	0.882	0.754
Qwen3-27B	0.877	0.655	0.693	0.795	0.685	0.375	0.752	0.689	0.820	0.736	0.681	0.850	0.740
GPT-5.4	0.868	0.738	0.802	0.740	0.675	0.500	0.708	0.662	0.831	0.653	0.638	0.853	0.738
GLM-4.6V	0.922	0.694	0.730	0.875	0.760	0.375	0.709	0.659	0.785	0.679	0.627	0.808	0.728
Grok-4.3	0.815	0.657	0.749	0.810	0.715	0.405	0.696	0.661	0.785	0.676	0.653	0.779	0.711
MiMo-V2.5	0.825	0.631	0.610	0.770	0.675	0.325	0.724	0.632	0.710	0.709	0.646	0.727	0.684
Llama-3.2-11B	0.856	0.505	0.486	0.765	0.655	0.245	0.794	0.676	0.610	0.782	0.648	0.706	0.671
GPT-5.4-mini	0.802	0.522	0.449	0.760	0.555	0.295	0.669	0.539	0.596	0.668	0.582	0.651	0.607

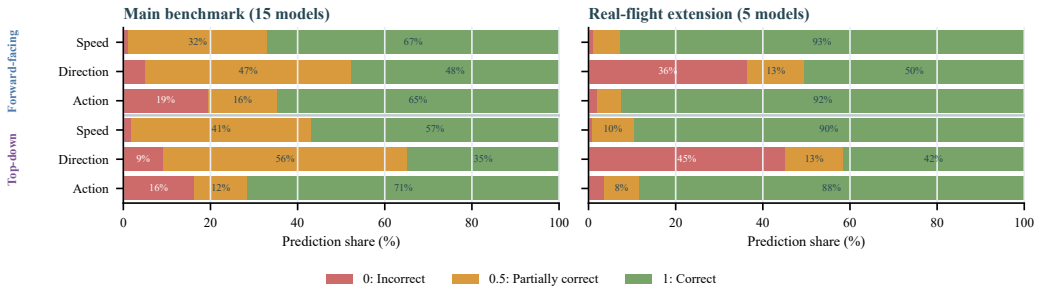


Figure 4: Component-score distributions by viewpoint on the curated benchmark (left) and the matched real-flight extension (right). Bars show the shares scored as incorrect (0), partial (0.5), or correct (1). Top-down views shift direction from fully correct toward partial or incorrect outcomes, and the same shift persists on onboard imagery while speed and action remain near ceiling.

Table 3: Results stratified by camera viewpoint. Model scores average 901 forward-facing and 449 top-down examples; Human averages 120 and 93 sampled examples, respectively. Total combines the three component scores. Δ Dir. is forward-facing Dir. minus top-down Dir.; positive values indicate lower top-down performance.

Model	Forward-Facing				Top-Down				Δ Dir.
	Speed↑	Dir.↑	Action↑	Total↑	Speed↑	Dir.↑	Action↑	Total↑	
Human	0.954	0.921	0.938	0.938	0.919	0.935	0.903	0.920	-0.015
GLM-5V-Turbo	**0.921**	**0.775**	0.820	**0.840**	**0.886**	**0.745**	**0.863**	**0.830**	+0.030
Mistral-Medium-3	0.916	0.719	**0.855**	0.829	0.880	0.618	0.845	0.778	+0.101
Qwen3.5-397B-A17B	0.862	0.767	0.791	0.807	0.841	0.722	0.820	0.793	+0.046
Gemma-4-31B	0.871	0.774	0.816	0.820	0.787	0.677	0.830	0.761	+0.097
Kimi-K2.6	0.853	0.724	0.774	0.784	0.786	0.668	0.841	0.761	+0.056
Qwen3.5-9B	0.837	0.768	0.725	0.779	0.820	0.685	0.780	0.760	+0.083
Doubao-Seed-1.6-Vision	0.875	0.741	0.670	0.767	0.870	0.663	0.746	0.760	+0.079
Gemini-2.5-Flash	0.837	0.720	0.754	0.771	0.765	0.616	0.787	0.719	+0.104
Qwen3-27B	0.812	0.720	0.750	0.761	0.737	0.588	0.778	0.697	+0.132
GPT-5.4	0.775	0.728	0.796	0.765	0.667	0.576	0.825	0.683	**+0.152**
GLM-4.6V	0.782	0.688	0.696	0.723	0.759	0.621	0.845	0.737	+0.067
Grok-4.3	0.759	0.684	0.721	0.721	0.680	0.617	0.790	0.691	+0.067
MiMo-V2.5	0.784	0.675	0.643	0.703	0.689	0.569	0.686	0.646	+0.106
Llama-3.2-11B-Vision	0.811	0.650	0.542	0.674	0.796	0.543	0.650	0.664	+0.106
GPT-5.4-mini	0.741	0.565	0.545	0.621	0.661	0.519	0.556	0.580	+0.046

ble 2 shows that scenario variation changes which component dominates the error, while direction remains the most consistent cross-model bottleneck.

Viewpoint-dependent direction errors. Forward-facing examples obtain 0.7577 overall, compared with 0.7240 for top-down examples, and most of the gap comes from direction. Direction falls from 0.7132 to 0.6284; its exact-match rate drops from 47.69% to 34.82%. Action instead increases from 0.7265 to 0.7761. Table 3 and Figure 4 show the top-down degradation to movement-frame grounding. Forward-facing views expose horizon, facade, and vanishing-point cues that connect target displacement to heading and depth. Top-down views are predominantly image-aligned and remove much of this projective evidence. A model may identify a target as above or below yet map that relation too directly to a command label. The frequent top-down substitutions from Down to None and from Right to Forward capture this ambiguity: correct spatial attention does not guarantee correct grounding in the aircraft movement frame.

Strict command accuracy. Only 25.07% of predictions match all command components exactly, compared with exact-match rates of 63.54% for speed, 43.41% for direction, and 66.93% for action. Across the 15 models, total score and full-command exact-match rate are strongly associated (Pearson $r = 0.944$; Spearman $\rho = 0.950$). The graded score therefore tracks strict command correctness while retaining information about near-miss behavior. Component intersections show how direction

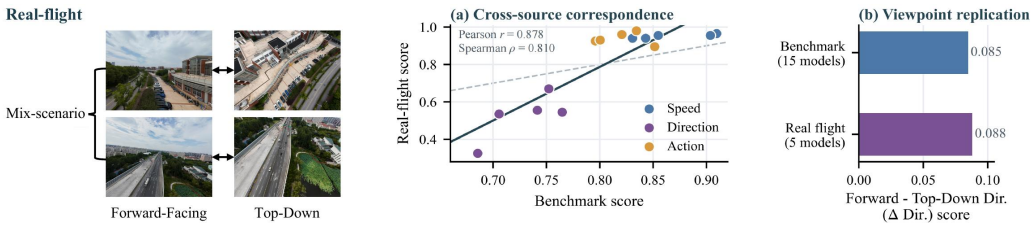


Figure 5: Validation in real flight. Component performance aligns with the curated benchmark across five models, while the forward-facing direction advantage persists on 50 matched onboard image pairs, indicating the movement-frame grounding bottleneck.

errors propagate into joint failure. In 17.08% of all evaluations, speed and action are both exactly correct but direction is not; the corresponding single-component blocker rates are 11.43% for action and 6.24% for speed. Even among predictions with the correct action, only 37.45% match the full command. Direction is thus the most frequent lone break in an otherwise correct command, explaining why relatively strong action recognition does not translate into reliable command selection.

6 Real-Flight Validation

6.1 Overall Real-Flight Performance

The real-flight extension checks whether the same component failure pattern holds on images captured by an operating drone. Across the five evaluated models, the mean total score is 0.7983. This aggregate conceals a component-specific pattern: speed reaches 0.9510 and action reaches 0.9380, whereas direction falls to 0.5260. Table 4 reports model-level results. Qwen3.5-397B-A17B ranks first at 0.8478, followed by GLM-5V-Turbo at 0.8225. Full-command exact match, which requires all three component scores to equal 1, averages 42%; model-level rates range from 23% to 55%. Model separation is concentrated in direction. Speed ranges only from 0.940 to 0.965, and action ranges from 0.895 to 0.980, while direction ranges from 0.325 for Mistral Medium 3 to 0.670 for Qwen3.5-397B-A17B. The same two models obtain the lowest and highest full-command exact-match rates, 23% and 55%, respectively. Thus, even when image capture and operation selection are reliable, resolving the movement axis determines whether the complete command succeeds.

6.2 Cross-Source Consistency

Figure 5 compares the five models evaluated on both the benchmark and real flights across speed, direction, and action. The 15 paired scores across models and components show strong cross-source correspondence (Pearson $r = 0.878$; Spearman $\rho = 0.810$), indicating that the component-level capability profile transfers from the benchmark to onboard imagery. On the benchmark these models obtain 0.8684 for speed, 0.7299 for direction, and 0.8206 for action; on real-flight images they obtain 0.9510, 0.5260, and 0.9380. Relative to the benchmark, speed rises by 0.0826 and action by 0.1174, while direction decreases by 0.2039. The real-flight domain therefore preserves the same weakest component while amplifying its separation from recognition-driven operation selection. Because speed and action remain high on both sources, the weighted total changes little even as direction falls by 0.2039. Comparing totals alone would therefore miss the movement-frame deficit that the component correspondence exposes. The scatter in Figure 5 makes this separation visible: speed and action occupy the high-score region for both sources, while direction forms the lower cluster. Pooling the five models, the 270 wrong real-flight directions fall into a few substitutions. Forward to Down (42) and Left to Down (40) are the most frequent, followed by Left to Forward (30) and Right to Forward (23); these four pairs alone cover 135 of the wrong predictions. As on the curated benchmark, the errors concentrate in a small set of direction pairs.

6.3 Matched-Viewpoint Analysis

The matched pairs sharpen the viewpoint comparison because each pair shares one task and action label while the required motion follows its image geometry. All 50 pairs retain the same action,

Table 4: Overall and matched-viewpoint results on 100 real-flight images in 50 matched task pairs (one forward-facing and one top-down image per pair, shared task). Overall: component scores, weighted Total, and full-command exact match (%). Δ Dir. is forward-facing minus top-down direction; positive values favor forward views. Boldface marks the best model result.

Model	Overall					Forward		Top-Down		Δ Dir.
	Speed↑	Dir.↑	Action↑	Total↑	Exact↑	Dir.↑	Total↑	Dir.↑	Total↑	
Qwen3.5-397B	0.955	**0.670**	0.930	**0.848**	**55.0%**	**0.750**	**0.890**	**0.590**	**0.806**	**+0.160**
GLM-5V-Turbo	**0.965**	0.545	**0.980**	0.823	50.0%	0.600	0.847	0.490	0.799	+0.110
Gemma-4-31B	0.940	0.555	0.960	0.811	43.0%	0.630	0.835	0.480	0.788	+0.150
Kimi-K2.6	0.940	0.535	0.925	0.794	39.0%	0.590	0.824	0.480	0.764	+0.110
Mistral-Medium-3	0.955	0.325	0.895	0.717	23.0%	0.280	0.707	0.370	0.726	-0.090
Mean	0.951	0.526	0.938	0.798	42.0%	0.570	0.820	0.482	0.776	+0.088

whereas 27 require a different direction and only three require a different speed. As shown in Table 4, direction decreases from 0.5700 on forward-facing images to 0.4820 on top-down images. Four of five models show a forward-facing advantage; Qwen3.5-397B-A17B and Gemma 4 31B show the largest gaps at 0.160 and 0.150, while Mistral Medium 3 reverses the aggregate trend by 0.090. Its reversal pairs the lowest forward-facing direction in Table 4 (0.280) with the lowest top-down direction (0.370), tracing the flipped sign to weak forward-facing grounding. The viewpoint change also affects joint command completion. Total score falls from 0.8204 to 0.7763 and full-command accuracy from 46% to 38%. The mean direction gap of 0.0880 nearly matches the 0.0848 gap in the 15-model benchmark, reproducing both the affected component and the magnitude of the viewpoint effect. Because the paired task and action remain fixed, the decline is specifically associated with interpreting image geometry as aircraft-centered movement.

The real-flight study therefore extends the benchmark result. Across curated and onboard imagery, models usually recover the task-relevant operation but remain less reliable when converting target relations into aircraft motion. This consistency motivates aerial agents that combine visual grounding with pose, heading, altitude, and camera information, or explicitly supervise viewpoint and coordinate conversion when those signals must be inferred.

7 Limitations and Responsible Use

Main-set labels and references are model-assisted; real-flight labels are manual but rationales are model-assisted. Matched pairs do not control pose, illumination or geometry, while scenario, viewpoint and command distributions are imbalanced; comparisons are diagnostic. DroneBench evaluates decision-layer command selection in isolation from low-level flight control. Coupling it with end-to-end systems that entangle perception, planning, and control, such as AerialVLA and ActiveFly-Bench (Huang et al., 2026; Deng et al., 2026), could extend the evaluation toward the full pipeline from decision to flight control. Endpoint and judge outputs are not deterministic, the judge requires human validation, and image release requires licensing.

8 Conclusion

To assess whether VLMs can translate aerial observations and result-oriented goals into aircraft-centered commands, we introduce DroneBench, a diagnostic benchmark for one-step command selection. Given an aerial image and a task, DroneBench requires VLMs to predict a structured command composed of speed, direction, and action. Across 1,350 examples and 15 models, models often select appropriate actions but remain less reliable at grounding movement direction, especially in top-down views. The same pattern appears in 50 matched forward-facing and top-down pairs collected during real flights, where direction remains much weaker than speed and action. These results identify viewpoint-aware spatial grounding as a central challenge for aerial VLMs and motivate aerial agents that carry explicit viewpoint information when converting perception into motion.

AI USE STATEMENT

In this work, we used generative AI tools for drafting and polishing prose, for retrieving related work during the literature survey, and for assisting dataset construction (viewpoint classification, task and rationale generation under human review). We have not used generative AI tools for establishing the research claims, for scoring model outputs (the fixed judge is a disclosed evaluation component applied uniformly to all systems), or for writing the reported experimental results. Additionally, we used generative AI tools for code assistance on evaluation infrastructure, which was verified by deterministic re-aggregation from released per-row records. We have reviewed all AI-assisted work, and we take responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

DroneBench is fully specified by released artifacts, namely the 1,350-row dataset (with embedded image payloads and reference commands), the complete per-example prediction and judgment records for all 20,250 main-benchmark and 500 real-flight evaluations, the canonical prompts and parsers, and deterministic aggregation scripts that recompute every reported number from per-row records. The test prompt is identical for all models and does not reveal the viewpoint-specific direction rules; scoring uses a fixed, cited judge model with a disclosed 0/0.5/1 rubric. The evaluation is inference-only, so there are no training hyperparameters. The technical appendix documents the real-flight acquisition protocol, pair design, leakage checks, and label distributions; released review samples include image payloads and checksums for verification.

REFERENCES

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, et al. RT-2: Vision-language-action models transfer web knowledge to robotic control, 2023.

ByteDance Seed Team. Introduction to techniques used in Seed1.6. https://seed.bytedance.com/en/blog/introducing-seed1-6, 2025.

Jianwei Chen, Yuhang Wang, Zhiting Liu, Feiyang Deng, Yiming Wang, Zhaoxiang Zhang, and Hao Zhao. Towards autonomous UAV visual object search in city space: Benchmark and agentic methodology, 2025.

Xiaoye Chen, Jiaming Wei, Ruiqi Li, Yujie Tan, Chenglin Xie, Jiangbo Lu, Xiaohua Zhou, Lin Lin, and Guanbin Zhu. OpenUAV: A UAV-need-help dataset for vision-language navigation in real-world environments, 2024.

Zhiyuan Deng, Yucheng Zhang, Chao Huang, Yixin Zhu, Siyuan Huang, and Hang Zhao. ActiveFly-Bench: Benchmarking embodied active perception with UAVs, 2026.

DJI. DJI Avata 2 user manual. DJI Download Center, https://www.dji.com/avata-2/downloads, 2024.

Danny Driess, Fei Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. PaLM-E: An embodied multimodal language model, 2023.

Dawei Du, Yuankai Qi, Hongyang Yu, Yifan Yang, Kaiwen Duan, Guorong Li, Weigang Zhang, Qingming Huang, and Qi Tian. The unmanned aerial vehicle benchmark: Object detection and tracking. In *European Conference on Computer Vision Workshops*, 2018.

Gemini Team. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities, 2025.

Gemma Team. Gemma 4 technical report, 2026.

GLM Team. GLM-5V: Generalizable, steerable, and enterprise-ready multimodal reasoning, 2026.

Yongchang Huang, Fan Huang, Xiang Li, Hanchen Hu, Heng Wang, Zihui Xue, Jun Wang, Anqi Cao, Yunfei Chen, Yuke Zhu, Hao Zhou, Alan Yuille, Chuang Chen, and Cewu Lu. AerialVLA: Towards embodied flight navigation with vision-language-action model, 2026.

Kartik Kuckreja, Muhammad Sohail Danish, Muzammal Naseer, Abhijit Das, Salman Khan, and Fahad Shahbaz Khan. GeoChat: Grounded large vision-language model for remote sensing, 2023.

Jungdae Lee, Hyo-Won Choi, Nagyeom Kim, Dae Young Noh, Taeyoung Kang, Tae-Han Moon, and Hyun Myung Kim. Language-goal aerial navigation dataset with geographic information, 2024.

Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. SEED-Bench: Benchmarking multimodal LLMs with generative comprehension, 2023.

Fangchen Liu, Wanting Han, Ziyu Liao, Lingfei Lu, Tao Chen, Xupeng Zhang, Xinzhu Ma, Quan Jiang, Yun Liu, Ling Shao, Yunhong Ding, Guang Zhan, and Dacheng Tao. AirCopBench: Benchmarking collaborative embodied intelligence in multi-drone systems, 2025.

Xuchen Liu, Jiawei Huang, Shihao Xia, Bingxi Liu, Jinqiang Cui, and Jiankun Yang. ImagineUAV: Aerial vision-language navigation via world-action modeling and kinodynamic planning, 2026.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-Eval: NLG evaluation using GPT-4 with better human alignment, 2023a.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. MMBench: Is your multi-modal model an all-around player?, 2023b.

Meta AI. Llama 3.2: Revolutionizing edge AI and vision with open, customizable models. https://ai.meta.com/blog/llama-3-2-connect-2024-vision-edge-mobile-devices/, 2024.

Mistral AI. Mistral medium 3. https://mistral.ai/news/mistral-medium-3, 2025.

Moonshot AI. Kimi K2.6. https://www.moonshot.ai/, 2026.

Dilxat Muhtar, Zhenshi Zhang, Peng Xiao, Zhi Li, and Feng Gu. LHRS-Bot: Empowering remote sensing with VGI-enhanced large multimodal language model, 2024.

OpenAI. GPT-4o system card, 2024.

OpenAI. Introducing GPT-5.4. https://openai.com/index/introducing-gpt-5-4/, 2026a.

OpenAI. Introducing GPT-5.4 mini and nano. https://openai.com/index/introducing-gpt-5-4-mini-and-nano/, 2026b.

Qwen Team. Qwen3 technical report, 2025.

Qwen Team. Qwen3.5: Towards native multimodal agents. https://github.com/QwenLM/Qwen3.5, 2026.

Shital Shah, Debadeepta Dey, Chris Lovett, and Ashish Kapoor. AirSim: High-fidelity visual and physical simulation for autonomous vehicles. In *Field and Service Robotics*, 2018.

Yunlong Song, Selim Naji, Elia Kaufmann, Antonio Loquercio, and Davide Scaramuzza. Flightmare: A flexible quadrotor simulator. In *Conference on Robot Learning*, 2020.

State Council of the People's Republic of China. Report on the work of the government. Government Work Report delivered at the Second Session of the 14th National People's Congress, 2024.

Ruipu Wu, Yige Zhang, Jinyu Chen, Linjiang Huang, Shifeng Zhang, Xu Zhou, Liang Wang, and Si Liu. AeroDuo: Aerial duo for UAV-based vision and language navigation, 2025.

xAI. Grok 4.3. https://x.ai/news/grok-4-3, 2026.

Xiaomi MiMo Team. MiMo-V2.5: Towards generalizable long-horizon agentic intelligence via rewarded meta-thinking and non-blocking tool use, 2026.

Shaoyuan Xie, Lingdong Kong, Yuhao Dong, Chonghao Sima, Wenwei Zhang, Qi Alfred Chen, Ziwei Liu, and Liang Pan. Are VLMs ready for autonomous driving? an empirical study from the reliability, data, and metric perspectives, 2025.

Fanglong Yao, Yuanchang Yue, Youzhi Liu, Xian Sun, and Kun Fu. AeroVerse: UAV-agent benchmark suite for simulating, pre-training, finetuning, and evaluating aerospace embodied world models, 2024.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. MM-Vet: Evaluating large multimodal models for integrated capabilities, 2024.

Z.ai. GLM-4.6V: Advanced visual reasoning with native tool use. https://github.com/zai-org/GLM-V, 2025.

Baining Zhao, Jiacheng Xu, Weicheng Feng, Xin Zhang, Zhaolu Wang, Haoyang Wang, Shilong Ji, Ziyou Wang, Jianjie Fang, Zhiheng Zheng, Weichen Zhang, Yu Shang, Wei Wu, Chen Gao, Xinlei Chen, and Yong Li. WorldVLN: Autoregressive world action model for aerial vision-language navigation, 2026.

Pengfei Zhu, Longyin Wen, Dawei Du, Xiao Bian, Huijuan Fan, Qinghua Hu, and Haibin Ling. Detection and tracking meet drones challenge. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11):7380–7399, 2022.