Motion-o: Trajectory-Grounded Video Reasoning

Bishoy Galoaa*
Northeastern University
galoaa.b@northeastern.edu

Shayda Moezzi*
Northeastern University
moezzi.s@northeastern.edu

Xiangyu Bai
Northeastern University
bai.xiang@northeastern.edu

Sarah Ostadabbas
Northeastern University
s.ostadabbas@northeastern.edu

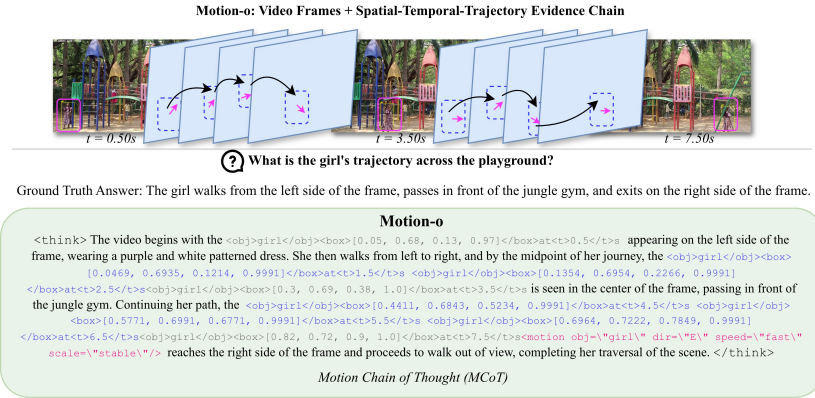


Figure 1: Although recent video models can produce fluent reasoning traces, their reasoning is typically ungrounded. Motion-o introduces a **Spatial–Temporal–Trajectory (STT) evidence chain** that explicitly embeds model reasoning with a trajectory-faithful trace.

Abstract

Recent video reasoning models increasingly produce spatio-temporal evidence chains that localize objects at specific timestamps. While these traces improve interpretability by grounding *where* and *when* evidence appears, they often leave the motion connecting observations, the *how*, implicit. This makes dynamic and trajectory-dependent claims difficult to supervise, verify, or penalize when unsupported by the video. We formalize this missing component as Spatial-Temporal-Trajectory (STT) reasoning and introduce **Motion-o**, a motion-centric extension to vision-language models (VLMs) that makes trajectories explicit and verifiable. Motion-o augments evidence chains with Motion Chain of Thought (MCoT), a structured pathway that represents object motion through a discrete `<motion/>` tag summarizing direction, speed, and scale change. To supervise MCoT, we densify sparse spatio-temporal annotations into object tracks and derive motion descriptors from centroid displacement and box-area change. We then train with complementary rewards for trajectory consistency and visual grounding, including a perturbation-based signal that penalizes motion descriptions that remain unchanged when temporal evidence is removed. Matched-density and training-stage controls separate the benefits of denser supervision, motion tags, and reinforcement learning. Temporal-order and predicted-box controls support motion-

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

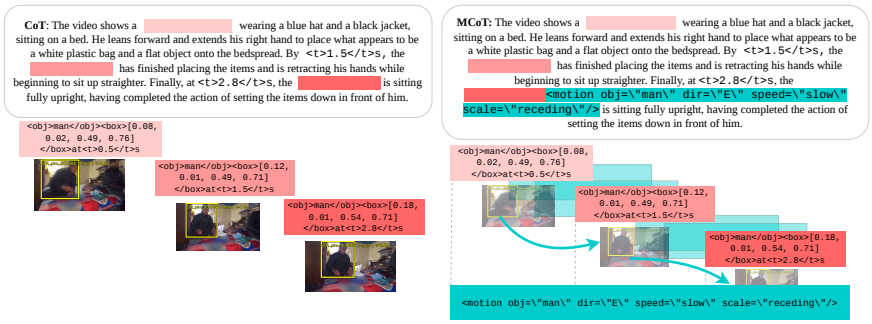


Figure 2: CoT vs. Motion Chain-of-Thought (MCoT). CoT in [16] yields sparse temporal bounding boxes, forcing implicit inter-frame interpolation. MCoT adds an object-conditioned `<motion/>` tag that parameterizes the dynamics between observations.

sensitive behavior, while oracle-track and egocentric evaluations expose limits of the coarse representation. These results support an explicit motion interface for evidence-based reasoning without architectural modifications. Code is available at `ostadabbas/Motion-o`.

1 Introduction

Humans effortlessly perceive motion. We recognize when a person walks leftward, when a ball accelerates, or when a vehicle follows a smooth arc. This capacity for motion understanding is so fundamental that it often goes unnoticed, yet it underpins our ability to predict, explain, and reason about the dynamic visual world. However, many current vision-language model (VLM) reasoning pipelines still represent evidence primarily as localized snapshots, and lack an explicit mechanism to reason about how objects move between observations. As a result, models frequently rely on textual interpolation to "fill in" dynamics, producing motion statements that are plausible but ungrounded and difficult to verify.

Recent work on grounded video reasoning has made progress toward structured visual evidence [6, 25, 16]. As illustrated in the left example in Figure 2, such traces improve interpretability by grounding objects in space and time. However, they remain snapshot-based: an `<obj>` and `<box>` at a few timestamps specify where an entity appears, but not the trajectory that connects those observations. The missing link is not motion as a low-level visual signal, which has been studied extensively in tracking, robotics, and video generation. Rather, the missing link is motion as an explicit and rewardable component of the model's reasoning trace. Without this component, a model may answer a trajectory-dependent question correctly or incorrectly while leaving no structured evidence for whether its motion claim follows from the visual dynamics [21, 29, 14].

We formalize this missing capability as **Spatial–Temporal–Trajectory (STT) reasoning**: the joint ability to reason about *where* objects are, *when* they appear, and *how* they move between observations. Existing evidence-based video reasoning frameworks largely address the spatial and temporal dimensions, but leave trajectory information implicit. This is limiting because video understanding requires explicit motion understanding: whether an object moved leftward, remained stationary, or changed apparent scale. Our current tag summarizes single-object motion in 2D image space; it does not encode curved paths or separate camera motion from object motion. Such descriptors provide the semantic link between discrete observations and dynamic interpretation, supporting motion-sensitive question answering, anomaly detection, trajectory prediction, and embodied reasoning.

We introduce **Motion-o**, a lightweight extension for making object motion explicit and verifiable in grounded video reasoning. Motion-o augments spatio-temporal evidence chains with **Motion Chain-of-Thought (MCoT)**, a structured reasoning step that summarizes the motion connecting grounded observations; figure 1 provides an overview of Motion-o and MCoT. After grounding an object at multiple timestamps, the model emits a compact self-closing `<motion/>` tag with discrete, interpretable attributes for direction, speed, and scale change; for example, `<motion obj="person" dir="W" speed="slow" scale="stable"/>`. This converts an implicit trajectory into a structured intermediate statement that can be parsed, supervised, rewarded, and inspected.

The central question of this work is not whether motion can be represented in principle. Prior work has explored visual traces for robotics, trajectory-based controls for generation, and spatio-temporal grounding for localization. Motion-o targets a different interface: open-ended video reasoning, where the model must connect localized observations into a trajectory-faithful explanation and final answer. The contribution is a reasoning format, training recipe, and dense-trajectory resource rather than a new architecture or motion vocabulary. The value of MCoT is that familiar motion primitives become structured, rewardable variables inside the VLM output. Controls in Section 4.2 separate the tag from denser supervision and test its dependence on temporal evidence.

To train Motion-o, we augment existing spatio-temporal annotations with denser object tracks and derive discrete motion descriptors from centroid displacement and box-area change. These descriptors teach the MCoT format during supervised fine-tuning, while trajectory and visual-grounding rewards encourage predicted motion tags to both match the observed track and depend on temporal visual evidence. Motion-o is intentionally simple and architecture-agnostic: it exposes object motion as a structured intermediate variable that can be supervised and verified within existing VLM pipelines.

Our contributions are as follows: (1) we formalize **Spatial–Temporal–Trajectory (STT) reasoning** as an extension of grounded video reasoning from localized snapshots to object motion between observations; (2) we introduce **Motion Chain-of-Thought (MCoT)**, a structured `<motion/>` reasoning operator that encodes motion as parseable intermediate evidence; (3) we propose **Motion-o**, an architecture-agnostic training framework that teaches and rewards explicit motion reasoning through supervised fine-tuning and trajectory-aware reinforcement learning; and (4) we construct a trajectory-grounding augmentation of existing spatio-temporal annotations, deriving dense object tracks and motion descriptors for supervising trajectory-faithful video reasoning.

2 Related Work

Spatial and Temporal Video Reasoning. Recent work has improved the ability of VLMs to ground video reasoning in explicit spatial and temporal evidence. Temporal grounding methods such as Time-R1 [26] and TVG-R1 [26] use verifiable rewards to identify when relevant evidence appears. Spatial grounding methods such as SpaceR [19] use reinforcement learning to improve object-centric localization. Open-o3 Video [16] emits timestamped frames and bounding boxes as structured video evidence, while STVG-o1 [9] uses box-level chain-of-thought for spatio-temporal grounding. These methods demonstrate the value of explicit evidence, but the inter-frame motion connecting localized observations remains implicit. Motion-o addresses this missing trajectory dimension by requiring the model to articulate how objects move between grounded observations.

Evidence-based Video Reasoning. A broader line of work studies explicit visual operations as intermediate reasoning steps, including detection, segmentation, tracking, and box-level grounding [5, 24, 31, 9]. DeepEyes [31] shows that reinforcement learning can incentivize image-tool reasoning, while TreeBench [23] provides methodology for traceable box-level evidence in images. In video, OpenAI-o3 [18] popularizes thinking with images, and Open-o3 Video [16] extends this idea to thinking with frames. Motion-o follows this evidence-based direction, but changes the evidence unit. Instead of only grounding objects at isolated frames, it makes the motion relation between grounded observations explicit.

Trajectory, Trace, and Action-Centric Representations. A related line of work represents motion through trajectories, traces, or action-conditioned structures. In robotics, visual-trace methods such as LLARVA [17] use 2D trace prediction to align perception with action, while spatio-temporal grounding models such as VideoMolmo [1] improve localization and pointing over time. These works show that explicit traces are useful interfaces for dynamic visual tasks. Motion-o is complementary, but rather than treating trajectory outputs as action targets or localization outputs, it inserts motion into the VLM reasoning trace as an explicit variable connecting grounded observations to the final response.

Reinforcement Learning-based training. Reinforcement learning has recently become a central tool for improving multimodal reasoning. Video-R1 [6] uses temporal-aware GRPO to improve video understanding, VideoChat-R1 [13] extends RL to spatio-temporal perception, Video-RTS [28] combines RL with test-time scaling, and DeepVideo-R1 [20] introduces difficulty-aware regularization for temporal structure. These methods show that RL can improve video reasoning, but primarily

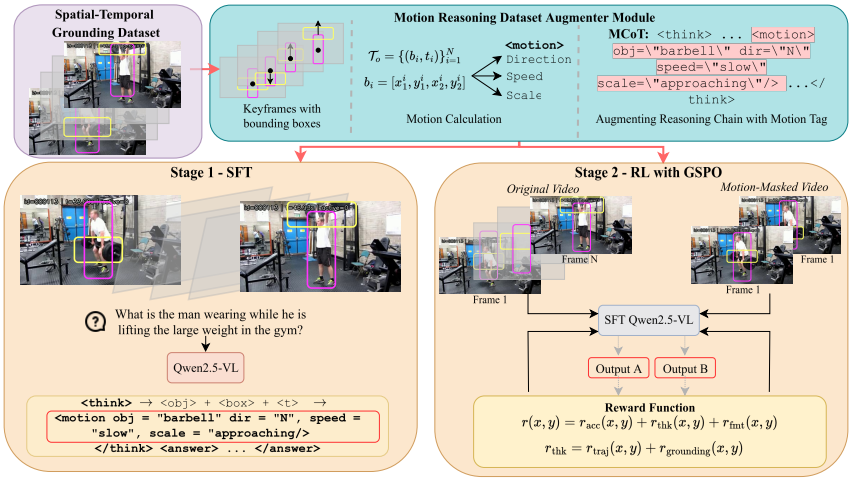


Figure 3: Motion-o pipeline. Starting from spatio-temporal grounding data [16] and our dense interpolated boxes, we derive motion primitives (direction, speed, scale) and insert them as MCoT `<motion/>` tags. Motion-o is then trained with SFT (Stage 1) to learn the STT/MCoT format and RL (GSPO, Stage 2) to reward trajectory-consistent tags that change when temporal evidence is removed.

optimize answer quality. Motion-o instead uses RL to supervise an explicit motion variable, where the `<motion/>` tag must agree with the trajectory and remain dependent on temporal visual evidence. Complementarily, Structure Over Scale [8] improves visual reasoning through GRPO on pedagogically structured videos.

Prior work has made substantial progress on temporal grounding, spatial grounding, and RL-based video reasoning. However, object motion is usually treated as an implicit latent dependency or an external control signal, rather than a verifiable statement inside the model's reasoning trace. Motion-o fills this gap by making trajectory information explicit and interpretable within grounded video reasoning.

3 Introducing Motion-o

We present **Motion-o**, a framework that extends grounded video reasoning with an explicit trajectory reasoning step (Figure 3). Motion-o augments spatio-temporal evidence chains with a discrete `<motion/>` descriptor that summarizes how an object moves between grounded observations. Given a video-question pair $(\mathcal{V}, q)$, the model outputs

$$y = \langle \texttt{think} \rangle R \langle / \texttt{think} \rangle \langle \texttt{answer} \rangle A \langle / \texttt{answer} \rangle,$$

where R contains (i) grounded spatio-temporal evidence and (ii) explicit motion descriptors. Motion-o is trained in two stages: (i) supervised fine-tuning (SFT) on motion-augmented reasoning traces that teach the model the structured format and syntax, and (ii) reinforcement learning that directly optimizes motion-centric rewards to encourage accurate and visually grounded motion descriptions.

Corresponding with Figure 3, we first summarize the spatio-temporal evidence format that we extend to STT evidence (Sec. 3.1), then introduce MCoT and how we obtain supervision for motion tags (Sec. 3.2), and finally describe the training pipeline and trajectory-grounded rewards (Sec. 3.3).

3.1 Spatial-Temporal-Trajectory (STT) Evidence Chains

Following Open-o3 Video [16], we adopt a structured format for grounding reasoning in visual evidence. Given a video $\mathcal{V}$ and question q, the model produces a response organized as $\langle$think$\rangle R \langle$/think$\rangle \langle$answer$\rangle A \langle$/answer$\rangle$, where R contains the reasoning trace and A the final answer. Within R, each grounded claim is written as:

$$\langle \text{obj} \rangle o \langle / \text{obj} \rangle \langle \text{box} \rangle b \langle / \text{box} \rangle \text{ at } \langle \text{t} \rangle t \langle / \text{t} \rangle \text{s}, \tag{1}$$

where o is the object name, $b = [x_1, y_1, x_2, y_2]$ is a normalized bounding box, and t is the timestamp in seconds. This representation creates an evidence chain that anchors language to concrete observations across time. For example, a model might write: "The man is visible holding a glass at 47.5s, and later at 54.2s is smiling," each claim anchored by a precise spatial location and timestamp. However, even when the same object is referenced at multiple timestamps, the evidence chain remains a set of discrete snapshots: it specifies where the object was at t_1 and t_2, but leaves the trajectory dynamics between t_1 and t_2 implicit. Motion-o targets precisely this missing link.

3.2 Motion Chain of Thought (MCoT)

We extend spatio-temporal evidence chains with explicit motion reasoning via a self-closing `<motion/>` tag to create STT evidence chains. The core idea is that after the model generates multiple temporal observations of the same object, it must summarize the motion that connects them. This converts a sequence of grounded boxes into a minimal, structured trajectory descriptor that is both interpretable and rewardable.

Motion Tag Format Each `<motion/>` tag carries discrete attributes: *obj* (object name), *dir* (direction: N, NE, E, SE, S, SW, W, NW, or STAT for stationary), *speed* (stationary, slow, moderate, fast), and *scale* (approaching, stable, receding). The format is:

$$\langle \texttt{motion obj} = \textit{name dir} = D \text{ speed} = S \text{ scale} = C \ /\rangle. \tag{2}$$

Unified Reasoning Flow. The complete reasoning chain then integrates temporal observations, spatial grounding, and motion reasoning into a single sequence. This enforces an explicit pathway: observe the object at multiple timestamps, ground each observation in a bounding box, emit a structured motion summary that links observations into a trajectory, and use that trajectory information to support the final answer. The `<motion/>` tag is a compact intermediate reasoning operator: it adds a small number of structured output tokens, but changes the semantics of the evidence chain by requiring the model to explicitly link grounded observations through a motion descriptor.

Automatic Motion Computation. To teach MCoT without additional human annotation, we generate discrete motion primitives directly from the bounding boxes provided in each sample in the STGR dataset [16]. Using the provided keyframe timestamps and object bounding boxes, we group bounding boxes by object identity across frames to obtain, for each object o, a time-ordered track $\mathcal{T}_o = \{(b_i, t_i)\}_{i=1}^{N}$, where $b_i = [x_1^i, y_1^i, x_2^i, y_2^i]$ denotes the box at time t_i. From $\mathcal{T}_o$, we compute box centroids c_i and areas a_i, and use the induced displacement sequence $\Delta c_i = c_{i+1} - c_i$ to summarize motion. We convert each track into a discrete motion descriptor with three attributes: **direction**, computed by aggregating per-step displacement directions over the track with jitter-reducing weights and mapping the result to compass bins plus a stationary state; **speed**, computed from the average displacement rate normalized by object scale and discretized into ordinal bins; and **scale change**, computed from the relative change in bounding-box area over time and discretized as {`approaching`, `stable`, `receding`}.

3.3 Training Pipeline

Our training follows the standard two-stage paradigm [16, 6] with minimal modifications, while introducing Motion-o as a format-level extension that explicitly supervises and rewards trajectory transition. We first teach the model to emit well-formed motion descriptors as part of the reasoning trace through SFT and then directly optimize motion correctness and visual grounding through RL.

Stage 1: Supervised Fine-Tuning. We fine-tune Qwen2.5-VL-7B [3] on the motion-augmented STGR data described above. Using standard cross-entropy optimization, the model learns to produce `<motion/>` tags as a structured continuation of spatio-temporal evidence chains: after grounding multiple observations of an object with `<obj><box><t>`, it emits a self-closing `<motion obj="..." dir="..." speed="..." scale="..."/>` that summarizes the trajectory connecting those observations. Beyond offline data augmentation, the only additional requirement in this stage is a system prompt instruction that specifies the tag schema and allowed attribute values. This stage establishes (i) strict format validity (`<think>/<answer>` and tag well-formedness), (ii) spatio-temporal grounding syntax, and (iii) the motion descriptor schema, ensuring the policy enters RL with stable, parseable trajectory reasoning traces.

Stage 2: Reinforcement Learning. We build upon the training recipe in [16], adopting Group Sequence Policy Optimization (GSPO) [30] to stably optimize long-form evidence-grounded video reasoning. Given a video–question prompt x, we sample a group of K candidate completions $\{y^{(k)}\}_{k=1}^{K} \sim \pi_\theta(\cdot \mid x)$, evaluate each completion with a scalar reward $r(x, y^{(k)})$ and compute a group-normalized advantage. GSPO then updates the policy using sequence-level importance ratios, which better aligns optimization with sequence-level rewards and improves stability for long-horizon reasoning.

Following [16], we decompose the scalar reward into three components:

$$r(x, y) = r_{\text{acc}}(x, y) + r_{\text{thk}}(x, y) + r_{\text{fmt}}(x, y). \tag{3}$$

where, r_{acc} encourages task-specific answer accuracy (e.g., exact match for MCQ, ROUGE for free-form QA), r_{fmt} enforces strict structured output formatting in the reasoning chain (including well-formed `<motion/>` tags), and r_{thk} incentivizes evidence-based reasoning. As in [16], we define the thinking reward as the sum of temporal and spatial grounding terms:

$$r_{\text{thk}}(x, y) \;=\; r_t(x, y) \;+\; r_s(x, y) \;+\; r_{\text{motion}}(x, y), \tag{4}$$

where r_t provides temporal supervision (with adaptive temporal proximity) and r_s provides spatial supervision with temporal gating to ensure spatial rewards are only computed when predicted timestamps are sufficiently close to ground-truth.

Motion-o extension: motion supervision inside r_{motion}. Our key extension is an explicit motion component

$$r_{\text{motion}}(x, y) \;=\; r_{\text{traj}}(x, y) \;+\; r_{\text{ground}}(x, y), \tag{5}$$

which teaches the model to (i) describe trajectory dynamics consistently with underlying tracks, and (ii) depend on temporal evidence. The coefficients of r_{acc}, r_{fmt}, r_t, r_s, r_{traj}, and r_{ground} are all 1.0, with no per-term scaling or tuning. Their natural score ranges need not be equal.

Trajectory Reward. For each object tracked across at least two temporal observations, we compute ground-truth motion bins from the bounding box trajectory and score the model's predicted `<motion/>` attributes via discrete bin matching with adjacency-aware partial credit:

$$r_{\text{traj}} = r_{\text{dir}} + r_{\text{speed}} + r_{\text{scale}}. \tag{6}$$

Ground-truth bins are derived from centroid displacement and box area dynamics: direction from the dominant displacement vector quantized to eight compass points plus stationary, speed from displacement magnitude normalized by the object's bounding box diagonal, and scale from the log-ratio of box areas between the first and last observation. Trajectory scores are computed per object identity to ensure consistency over a single track.

Visual Grounding Reward (Dual-Chain Verification). To encourage the model to derive motion descriptions from actual visual observation rather than textual interpolation, we introduce a visual grounding reward via dual-chain verification (see Stage 2 inputs in Figure 3). For each sample, we generate a second reasoning chain from a *motion-masked* video: only spans associated with motion tags are frozen using repeated keyframe copies, preserving appearance and leaving the rest of the video intact. Both chains use the same unified system prompt (Appendix F). We then compare the `<motion/>` tags produced from the original video against those from the motion-masked input:

$$r_{\text{ground}} \;=\; \mathbf{1}[d \neq d'] \;+\; \mathbf{1}[s \neq s'] \;+\; \mathbf{1}[c \neq c'], \tag{7}$$

where d, s, c and d', s', c' are the direction, speed, and scale attributes from the original and masked outputs. The sum ranges from 0 to 3 and credits attribute changes; it does not directly grade the correctness of the masked completion. In the reported recipe, an object absent from the masked output receives full grounding credit. This permits a dropping-object shortcut in principle. Masked-chain logs are predominantly stationary (81.4%), while absent objects account for 7.3% (Appendix H). Assigning zero credit to absent objects is a straightforward safeguard, but that variant is not the basis of the reported results.

Complementary Roles. The trajectory reward scores correctness against track-derived labels, whereas the perturbation reward supplies a proxy for dependence on temporal evidence. Neither proxy alone guarantees motion understanding. We therefore evaluate matched-density boxes-only controls, temporal perturbations, and oracle tracks in addition to aggregate benchmark scores (Section 4.2).

Table 1: Performance on the **V-STAR** benchmark, which evaluates spatio-temporal reasoning across three dimensions. Chain1 denotes *what–when–where*; Chain2 corresponds to *what–where–when*. mAM is the average of arithmetic mean, and mLGM is the average of modified logarithmic geometric mean, combining temporal and spatial alignment. * indicates re-evaluation using the vLLM framework with 16 sampled frames. The r_{ground} column indicates whether the visual grounding reward is enabled (👁) or disabled (🚫) in our methods. **Bold** numbers denote the best results, while underlined numbers indicate the second best.

Model	r_{ground}	What	When (Temporal IoU)		Where (Visual IoU)		Overall	
		Acc	Chain1	Chain2	Chain1	Chain2	mAM	mLGM
GPT-4o		60.8	16.7	12.8	6.5	3.0	26.8	38.2
Gemini-2-Flash		53.0	24.5	23.8	4.6	2.2	26.9	35.6
Video-LLaMA3		41.9	23.0	23.1	0.9	0.2	21.7	27.0
LLaVA-Video		49.5	10.5	12.2	1.9	1.3	20.8	27.3
VideoChat2		36.2	13.7	12.5	2.5	1.0	17.0	20.3
Oryx-1.5-7B		20.5	13.5	14.8	10.1	3.5	15.1	13.8
InternVL-2.5-8B		44.2	8.7	7.8	0.7	0.1	17.6	24.9
Qwen2.5-VL-7B* (base)		33.5	15.4	13.8	17.0	2.5	19.3	22.4
Qwen3-VL-8B* (base)		36.0	23.5	13.9	8.5	6.8	20.2	23.1
TRACE		17.6	19.1	17.1	0.0	0.0	12.0	13.3
Sa2VA-8B		16.4	0.1	0.0	32.3	37.5	17.1	20.3
Open-o3 Video		61.0	24.5	24.0	25.4	6.0	33.7	46.6
Motion-o (Qwen2.5-VL-7B)	🚫	62.2	26.2	25.8	26.6	12.3	35.0	48.8
Motion-o (Qwen2.5-VL-7B)	👁	62.7	26.6	26.2	28.2	15.1	35.5	49.4
Motion-o (Qwen3-VL-8B)	🚫	63.1	26.8	26.4	29.5	16.8	35.7	49.6
Motion-o (Qwen3-VL-8B)	👁	63.6	27.0	26.7	31.2	18.4	36.1	50.1
Open-o3 + MCoT	🚫	63.4	26.3	25.6	30.5	31.4	35.2	49.1
Open-o3 + MCoT	👁	**64.1**	**27.3**	**26.8**	**33.6**	**38.1**	**36.6**	**50.5**

3.4 Dataset Augmentation: Trajectory Grounding with Dense Motion Annotations

Existing spatio-temporal grounding datasets often provide bounding boxes at a small set of timestamps, yielding sparse trajectories that under-specify the motion connecting observations [16]. We therefore construct a trajectory-grounding augmentation of the PerceptionLM (PLM) subset [4] by converting each sparse set of timestamped boxes into a temporally denser trajectory. For each tracked object, we preserve the original annotated keyframes and insert intermediate supervision points between adjacent timestamps using the dense mask annotations available in PLM. This produces a denser sequence of (t, b_t) pairs for each tracked object, from which we derive the direction, speed, and scale-change descriptors described above. We release the augmented trajectories, derived motion descriptors, and splits to support reproducibility and to enable future work on explicit motion reasoning. The resource is intended primarily as a trajectory-grounding augmentation for evidence-based video reasoning. Further details on dataset splits and distribution are provided in the appendix in Section A and Figure 4.

4 Experimental Results

We train Motion-o from Qwen2.5-VL-7B [3] following the two-stage pipeline in Sec. 3.3. To test whether MCoT provides additive gains on an already strong evidence-based model, we also add MCoT to the Open-o3 Video [16] checkpoint (denoted *Open-o3 + MCoT*). Both variants are evaluated with and without the visual grounding reward r_{ground}. Following Open-o3 Video [16], we report on the **V-STAR** benchmark [27] for spatio-temporal reasoning (What / When / Where, mAM, mLGM), and on **VideoMME** [7], **WorldSense** [10], **TVGBench** [26], **MVBench**[12], and **MotionBench**[11] for general video understanding.

4.1 Main Results

Quantitative Analysis. As reported in Table 1 on the **V-STAR** benchmark, Open-o3 + MCoT achieves 36.6 mAM and 50.5 mLGM, improving over Open-o3 Video by +2.9 / +3.9 points. The gains are broad: What accuracy rises from 61.0 to 64.1, temporal grounding improves on both chains (Chain1: 24.5 $\rightarrow$ 27.3, Chain2: 24.0 $\rightarrow$ 26.8), and spatial grounding improves substantially (Chain1: 25.4 $\rightarrow$ 33.6, Chain2: 6.0 $\rightarrow$ 38.1), surpassing even the specialist Sa2VA-8B model on Where. Motion-o, initialized from Qwen2.5-VL-7B, reaches 35.5 mAM, already surpassing Open-o3 Video and all open-source baselines. Removing r_{ground} reduces mAM by 0.5 for Motion-o and 1.4 for Open-o3 + MCoT, while the motion-centric effect is larger: for Motion-o, it reduces MotionBench by 5.0 points and tag accuracy by 6.5. The stage decomposition in Table 4 shows that SFT provides much of the initial V-STAR gain, while RL further improves motion-centric performance. Beyond answer-level benchmark accuracy, we also evaluate whether the generated `<motion/>` tags match the ground truth, trajectory-derived labels in our augmented dataset (results in Appendix B and Table S6). Motion-o produces tags that are aligned with the ground truth tracks, and removing the visual grounding reward reduces this alignment. This supports our claim that MCoT is not only a readable reasoning format, but also a measurable intermediate representation tied to trajectory evidence.

In Table 2, Open-o3 + MCoT reaches 69.7 on VideoMME Overall and 41.5 on WorldSense, improving over Open-o3 Video by +6.1 and +4.0 points respectively while narrowing the gap with GPT-4o. Motion-o similarly improves over Qwen2.5-VL-7B across all metrics (+5.1 VideoMME, +3.1 WorldSense). These results show improvements on several video-understanding measures; they do not establish uniform gains on every task. Both Tables 1 and 2 compare similar-scale Qwen2.5 and Qwen3 backbones. The modest difference motivates examining localization, but does not establish it as the only bottleneck. The oracle-track control raises tag accuracy from 85.2% to 96.8% and fine-motion subset accuracy from 62.4% to 84.5% (Appendix H.1), supporting localization as an important constraint while leaving a residual gap. Separately, emitted tags outperform deterministic descriptors computed from the model's predicted boxes (Table 5), so the tag is not simply an arithmetic readout of the emitted coordinates.

MCoT also need not lengthen grounded reasoning: Motion-o and Open-o3 + MCoT use 19% and 28% fewer response tokens than Open-o3 Video, respectively (Appendix D).

Qualitative Analysis. Figure 4 shows two representative outputs. In the first example, the model tracks Sheldon across three timestamps with varying camera angles and emits two `<motion/>` tags, both correctly identifying the subject as stationary despite significant viewpoint changes. The dense multi-point grounding enables the model to distinguish true stationarity from apparent visual displacement caused by camera cuts. Stationary tags make a falsifiable claim that an object's position and scale remain stable across the evidence window. Appendix E provides an additional example where a `STAT` tag directly supports the final answer by verifying that the relevant background object persists across time. In the second example, the model grounds a duck at four consecutive timestamps and summarizes its eastward trajectory with a single motion tag. Additional qualitative examples and analyses are provided in the appendix and supplementary.

4.2 Ablation Studies

Separating Annotation Density and the Motion Tag. Table 3 crosses sparse/dense supervision with tag/no-tag reasoning. Denser boxes improve V-STAR even without the tag (28.0 to 33.2 mAM). At fixed dense supervision, adding the tag further improves mAM by 2.3 and mLGM by 2.9; in the sparse setting it adds 3.2 mAM. Both data density and the representation therefore contribute. VideoMME is not monotonic in density: sparse tags reach 68.1 versus 67.5 for dense tags. The numerical-tag variant performs poorly under this recipe (2.1 mAM), supporting the discrete choice here rather than proving that continuous representations cannot work.

Separating SFT, RL, and the Grounding Reward. Table 4 holds the initialization and data pipeline fixed. SFT establishes the output format and raises mAM from 19.3 to 32.5. Full RL adds 3.0 mAM, 8.0 MotionBench points, and 21.6 points of tag accuracy over SFT. Within RL, adding r_{ground} contributes 0.5 mAM, 5.0 MotionBench points, and 6.5 tag-accuracy points. Its practical impact is concentrated on motion-centric measures rather than aggregate V-STAR alone. Motion-o and

Table 2: Performance across video understanding, temporal grounding, and motion-centric benchmarks. †EgoSchema uses the 500-question MCQ subset; dashes indicate unreported results. The r_{ground} column indicates whether the visual grounding reward is enabled (👁) or disabled (🚫) in our methods. **Bold** numbers denote the best results, while underlined numbers indicate the second best.

Model	r_{ground}	VideoMME		WorldSense		TVGBench	MVBench	MotionBench	EgoSchema†
		Overall	Long	Overall	Recognition				
GPT-4o		71.9	-	**42.6**	-	26.8	43.50	33.0	–
Gemini-2.5-Pro		84.8	-	65.1	-	30.3	45.5	66.3	–
Gemini-3.0		**86.9**	-	-	-	33.6	65.5	70.4	–
Qwen2.5-VL-7B (base)		62.4	50.8	36.1	33.7	16.3	66.9	35.0	65.2
Qwen3-VL-8B (base)		64.1	52.3	37.2	35.1	18.4	69.0	38.0	–
VideoRFT-7B		59.8	50.7	38.2	36.6	14.3	61.4	36.2	–
VideoR1-7B		61.4	50.6	35.5	32.8	9.6	62.7	38.5	–
Open-o3 Video		63.6	54.9	37.5	36.8	20.8	64.4	45.0	56.8
Motion-o (Qwen2.5-VL-7B)	🚫	64.2	55.5	37.8	37.1	35.6	65.2	55.0	–
Motion-o (Qwen2.5-VL-7B)	👁	67.5	58.6	39.2	38.5	37.6	65.8	60.0	62.3
Motion-o (Qwen3-VL-8B)	🚫	64.8	56.1	38.4	37.7	36.8	66.1	57.0	–
Motion-o (Qwen3-VL-8B)	👁	68.2	59.3	40.1	39.4	38.9	67.0	62.0	–
Open-o3 + MCoT	🚫	66.3	57.3	38.9	38.2	38.2	67.2	60.0	–
Open-o3 + MCoT	👁	69.7	**60.3**	41.5	**40.8**	**39.6**	**69.2**	**63.0**	**67.4**



Question: Where is the board that Sheldon is holding while he is singing?
Reasoning: <think>The video shows a scene where four men are sitting on a couch, engaged in a game. Sheldon in the blue t-shirt is holding a clipboard and a pencil, which he uses to write down answers. This action is clearly visible throughout the segment. At one point, the camera focuses on Sheldon, who is seen writing on his clipboard. The setting is a living room decorated for Christmas, with a tree and lights visible in the background.
At the beginning of the video, <obj>Sheldon</obj><box>[198,10,440,336]</box>at<t>0</t>s is seated on the couch, holding a clipboard and a pencil, and appears to be writing something. <obj>Sheldon</obj><box>[198,10,440,336]</box>at<t>30</t>s<motion obj="Sheldon" dir="STAT" speed="stationary" scale="stable"/> continues this action, as he is shown writing on the clipboard again. <obj>Sheldon</obj><box>[198,10,440,336]</box>at<t>50</t>s<motion obj="Sheldon" dir="STAT" speed="stationary" scale="stable"/> is still holding the clipboard and pencil, confirming that he is actively using the board during the game.</think>
Answer: The board that Sheldon is holding is a clipboard. He is using a yellow pencil to write on it while participating in the game.



Question: Where is the duck behind another duck in the water between 16.0s~21.0s?
Reasoning: <think>The video shows three ducks swimming together. At 16.0s, the <obj>duck</obj><box>[160,100,224,140]</box>at<t>16.0</t>s is positioned behind the other two ducks. As they swim forward, the <obj>duck</obj><box>[160,100,224,140]</box>at<t>18.0</t>s remains consistently behind the others. By 19.0s, the <obj>duck</obj><box>[160,100,224,140]</box>at<t>19.0</t>s is still behind the group. This pattern continues until 21.0s, where the <obj>duck</obj><box>[160,100,224,140]</box>at<t>21.0</t>s<motion obj="duck" dir="E" speed="moderate" scale="stable"/> is clearly seen behind the other two ducks as they move through the water.</think>
Answer: The duck is behind another duck in the middle of the group, swimming towards the right side of the frame.

Figure 4: Qualitative examples of Motion-o reasoning. **Top:** Motion-o tracks Sheldon across camera changes and correctly emits stationary <motion/> tags, distinguishing true stationarity from apparent displacement. **Bottom:** Motion-o grounds a swimming duck across timestamps and summarizes its eastward trajectory with a single moderate-speed motion tag.

Open-o3 + MCoT use the same 34K RL corpus but different SFT mixtures and starting checkpoints (Appendix A); comparisons between those variants do not isolate MCoT by themselves.

Temporal Evidence and Predicted-Box Arithmetic. Emitted tags exceed deterministic descriptors computed from the same model's predicted boxes by 15.6 macro-average points (Table 5(a)). Under reversal, direction predictions invert by 180° in 88.2% of cases, and direction accuracy against inverted labels remains 74.5%; speed changes little (88.3% to 87.1%). Shuffling substantially reduces direction and speed accuracy (panel b). These interventions support dependence on temporal order; the box-arithmetic control separately shows that the emitted coordinates alone do not reproduce tag accuracy. They do not uniquely identify the internal mechanism or resolve the limits of the coarse schema.

Table 3: Density × tag control on V-STAR and VideoMME. The matched-density rows isolate the tag's marginal contribution. The last row tests a numerical-tag variant under the existing recipe.

Annotation density	Reasoning format	mAM	mLGM	VideoMME
Sparse	Boxes only, no tag	28.0	40.0	64.0
Sparse	Discrete tag	31.2	44.5	68.1
Dense	Boxes only, no tag	33.2	46.5	66.8
Dense	Discrete tag (default)	35.5	49.4	67.5
Dense	Numerical tag	2.1	5.4	14.4

Table 4: Training-stage decomposition for the Qwen2.5-VL-7B initialization. Tag accuracy is the macro-average of direction, speed, and scale exact matches. The base model's 0.0 reflects the absence of the prescribed tag, not zero motion perception.

Training stage	V-STAR mAM	MotionBench	Tag acc. (%)
Base	19.3	35.0	0.0
SFT (dense data + tags)	32.5	52.0	63.6
SFT + RL, no r_{ground}	35.0	55.0	78.7
SFT + RL, full	35.5	60.0	85.2

Table 5: Motion-o (Qwen2.5-VL-7B) controls (%). **(a)** Emitted attributes versus deterministic descriptors from the model's own predicted boxes. **(b)** Temporal-order interventions; reversed-video direction is scored against inverted labels. Inversion is the share of predictions changing by $180°$ relative to canonical video.

(a) Tag versus box arithmetic

Attribute	Tag	Boxes	Δ
Direction	78.9	64.2	+14.7
Speed	88.3	68.5	+19.8
Scale	88.4	76.1	+12.3
Macro avg.	85.2	69.6	+15.6

(b) Temporal order

Input	Direction	Speed	Inversion
Canonical	78.9	88.3	–
Reversed	74.5	87.1	88.2
Shuffled	18.2	22.0	–

Scope Beyond the Default Setting. On the 500-question EgoSchema [15] probe (Table 2), Motion-o scores 62.3 versus 56.8 for Open-o3 Video, but remains below the Qwen2.5 base (65.2); Open-o3 + MCoT reaches 67.4. Transfer is therefore useful but incomplete. A small distance-based grounding-reward pilot also improves over omitting that reward while retaining discrete tags (Appendix H.3); it does not test a fully continuous motion representation. Zero-shot schema prompting remains below fine-tuned performance (Appendix C).

5 Conclusion

We introduced Motion-o, a lightweight framework that makes object motion explicit and verifiable in grounded video reasoning. By augmenting spatio-temporal evidence chains with MCoT `<motion/>` tags, Motion-o connects *where* and *when* objects appear with *how* they move. Matched-density and training-stage controls support distinct contributions from denser supervision, motion tags, and RL. Temporal interventions and tag-versus-box comparisons support motion-sensitive behavior, while the coarse schema and imperfect grounding reward bound the claim. Future work will extend MCoT beyond coarse single-object motion toward richer interaction, occlusion, camera-motion, and tracking-aware representations.

Funding and Competing Interests

This project received no funding. The authors declare no competing financial interests.

References

[1] Ghazi Shazan Ahmad, Ahmed Heakl, Hanan Gani, Abdelrahman Shaker, Zhiqiang Shen, Ranjay Krishna, Fahad Shahbaz Khan, and Salman Khan. Videomolmo: Spatio-temporal grounding meets pointing, 2025. URL `https://arxiv.org/abs/2506.05336`.

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

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

[4] Jang Hyun Cho, Andrea Madotto, Effrosyni Mavroudi, Triantafyllos Afouras, Tushar Nagarajan, Muhammad Maaz, Yale Song, Tengyu Ma, Shuming Hu, Suyog Jain, et al. Perceptionlm: Open-access data and models for detailed visual understanding. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[5] Yue Fan, Xuehai He, Diji Yang, Kaizhi Zheng, Ching-Chen Kuo, Yuting Zheng, Xinze Guan, and Xin Eric Wang. Grit: Teaching mllms to think with images. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[6] Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, Junfei Wu, Xiaoying Zhang, Benyou Wang, and Xiangyu Yue. Video-r1: Reinforcing video reasoning in mllms. *arXiv preprint arXiv:2503.21776*, 2025.

[7] Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24108–24118, 2025.

[8] Bishoy Galoaa, Xiangyu Bai, and Sarah Ostadabbas. Structure over scale: Learning visual reasoning from pedagogical video. *arXiv preprint arXiv:2601.23251*, 2026. URL `https://arxiv.org/abs/2601.23251`.

[9] Xin Gu, Haoji Zhang, Qihang Fan, Jingxuan Niu, Zhipeng Zhang, Libo Zhang, Guang Chen, Fan Chen, Longyin Wen, and Sijie Zhu. Thinking with bounding boxes: Enhancing spatio-temporal video grounding via reinforcement fine-tuning. *arXiv preprint arXiv:2511.21375*, 2025.

[10] Jack Hong, Shilin Yan, Jiayin Cai, Xiaolong Jiang, Yao Hu, and Weidi Xie. Worldsense: Evaluating real-world omnimodal understanding for multimodal llms. *arXiv preprint arXiv:2502.04326*, 2025.

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

[12] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22195–22206, 2024.

[13] Xinhao Li, Ziang Yan, Desen Meng, Lu Dong, Xiangyu Zeng, Yinan He, Yali Wang, Yu Qiao, Yi Wang, and Limin Wang. Videochat-r1: Enhancing spatio-temporal perception via reinforcement fine-tuning. *arXiv preprint arXiv:2504.06958*, 2025.

[14] Yuanxin Liu, Shicheng Li, Yi Liu, Yuxiang Wang, Shuhuai Ren, Lei Li, Sishuo Chen, Xu Sun, and Lu Hou. Tempcompass: Do video llms really understand videos? *arXiv preprint arXiv:2403.00476*, 2024.

[15] Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. EgoSchema: A diagnostic benchmark for very long-form video language understanding. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://arxiv.org/abs/2308.09126.

[16] Jiahao Meng, Xiangtai Li, Haochen Wang, Yue Tan, Tao Zhang, Lingdong Kong, Yunhai Tong, Anran Wang, Zhiyang Teng, Yujing Wang, et al. Open-o3 video: Grounded video reasoning with explicit spatio-temporal evidence. *arXiv preprint arXiv:2510.20579*, 2025.

[17] Dantong Niu, Yuvan Sharma, Giscard Biamby, Jerome Quenum, Yutong Bai, Baifeng Shi, Trevor Darrell, and Roei Herzig. Llarva: Vision-action instruction tuning enhances robot learning. In Pulkit Agrawal, Oliver Kroemer, and Wolfram Burgard, editors, *Proceedings of The 8th Conference on Robot Learning*, volume 270 of *Proceedings of Machine Learning Research*, pages 3333–3355. PMLR, 06–09 Nov 2025. URL https://proceedings.mlr.press/v270/niu25a.html.

[18] OpenAI. Introducing o3 and o4-mini. https://openai.com/index/introducing-o3-and-o4-mini/, 2025.

[19] Kun Ouyang, Yuanxin Liu, Haoning Wu, Yi Liu, Hao Zhou, Jie Zhou, Fandong Meng, and Xu Sun. Spacer: Reinforcing mllms in video spatial reasoning. *arXiv preprint arXiv:2504.01805*, 2025.

[20] Jinyoung Park, Jeehye Na, Jinyoung Kim, and Hyunwoo J Kim. Deepvideo-r1: Video reinforcement fine-tuning via difficulty-aware regressive grpo. In *NeurIPS*, 2025.

[21] Maya Varma, Jean-Benoit Delbrouck, Sophie Ostmeier, Akshay S Chaudhari, and Curtis Langlotz. TRove: Discovering error-inducing static feature biases in temporal vision-language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[22] Leandro von Werra, Younes Belkada, Lewis Tunstall, Edward Beeching, Tristan Thrush, Nathan Lambert, Shengyi Huang, Kashif Rasul, and Quentin Gallouédec. TRL: Transformer reinforcement learning. https://github.com/huggingface/trl, 2020.

[23] Haochen Wang, Xiangtai Li, Zilong Huang, Anran Wang, Jiacong Wang, Tao Zhang, Sule Bai, Zijian Kang, Jiashi Feng, Wang Zhuochen, et al. Traceable evidence enhanced visual grounded reasoning: Evaluation and method. In *The Fourteenth International Conference on Learning Representations*, 2026.

[24] Jiacong Wang, Zijian Kang, Haochen Wang, Haiyong Jiang, Jiawen Li, Bohong Wu, Ya Wang, Jiao Ran, Xiao Liang, Chao Feng, et al. Vgr: Visual grounded reasoning. *arXiv preprint arXiv:2506.11991*, 2025.

[25] Qi Wang, Yanrui Yu, Ye Yuan, Rui Mao, and Tianfei Zhou. Videorft: Incentivizing video reasoning capability in mllms via reinforced fine-tuning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[26] Ye Wang, Ziheng Wang, Boshen Xu, Yang Du, Kejun Lin, Zihan Xiao, Zihao Yue, Jianzhong Ju, Liang Zhang, Dingyi Yang, et al. Time-r1: Post-training large vision language model for temporal video grounding. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[27] Yuxuan Wang, Zilong Zheng, Xueliang Zhao, Jinpeng Li, Yueqian Wang, and Dongyan Zhao. Vstar: A video-grounded dialogue dataset for situated semantic understanding with scene and topic transitions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5036–5048, 2023.

[28] Ziyang Wang, Jaehong Yoon, Shoubin Yu, Md Mohaiminul Islam, Gedas Bertasius, and Mohit Bansal. Video-rts: Rethinking reinforcement learning and test-time scaling for efficient and enhanced video reasoning. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 28114–28128, 2025.

[29] Xiantao Zhang. The escalator problem: Identifying implicit motion blindness in ai for accessibility. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 6635–6643, October 2025.

[30] Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, Rui Men, An Yang, et al. Group sequence policy optimization. *arXiv preprint arXiv:2507.18071*, 2025.

[31] Ziwei Zheng, Michael Yang, Jack Hong, Chenxiao Zhao, Guohai Xu, Le Yang, Chao Shen, and Xing Yu. Deepeyes: Incentivizing thinking with images via reinforcement learning. *arXiv preprint arXiv:2505.14362*, 2025.

Appendix

This appendix provides additional details supporting the methodology, evaluation, and reproducibility of Motion-o. We first describe the training data, implementation setup, and compute resources used for the SFT and RL stages. We then report motion-tag accuracy on the augmented trajectory dataset, including zero-shot `<motion/>` prompting controls, and provide extended benchmark results and ablations. Finally, we document the prompt templates used for training, inference, and evaluation, and discuss broader impacts and responsible-use considerations.

A Training Details

A.1 Training Data

Both Motion-o and Open-o3 + MCoT models reported in this paper are trained using data derived from the Spatio-Temporal Grounded Reasoning (STGR) dataset introduced in the Open-o3 paper [16]. Following that formulation, our training data is organized into SFT and RL stages, with each sample associated with a source dataset and a task type. The two models in this work, share the same RL data, but differ in their SFT mixtures to reflect different training priorities.

Shared RL data. For both Motion-o and Open-o3 + MCoT, the RL stage uses the same file, containing 34k training samples. This shared RL set includes five task categories: *temporal-spatial free-form QA* (12,047), *General video QA MCQ* (12,998), *visual QA* (5,000), *temporal QA* (2,279), and *general video QA free-form* (2,000). Using the same RL corpus for both models ensures that differences in final behavior are primarily attributable to the SFT data design rather than to different reinforcement learning supervision.

Motion-o SFT data. For Motion-o, the SFT stage uses a subset of the STGR dataset containing 22k samples. This mixture is broader and more balanced, combining motion-relevant grounded examples with a substantial amount of general video QA. Its task composition is: *general video QA free-form* (2,000), *General video QA MCQ* (13,000), *temporal QA* (491), and *temporal-spatial free-form QA* (7,047). This design preserves broad video reasoning coverage while still retaining a strong spatio-temporal grounding component, allowing Motion-o to couple explicit motion reasoning with more general video understanding capacity. Figure S5 summarizes the distribution of motion labels in our augmented dataset. Examples are not dominated by a single direction: eastward and westward motions are closely matched, and the remaining directional bins are also populated across the compass. This indicates that the learned `<motion/>` tags are trained on a reasonably balanced set of trajectory orientations, rather than reflecting a strong directional skew. The inner and outer rings further show that the data covers a range of speed and scale-change categories, providing supervision over multiple facets of motion behavior.

Open-o3 + MCoT SFT data. For Open-o3 + MCoT, we use the more motion-focused SFT mixture containing 7,838 samples. Its task composition is heavily concentrated in grounded motion-centric supervision: *temporal-spatial free-form QA* (6,849), *general video QA MCQ* (817), *general video QA free-form* (142), and *temporal QA* (30). We use this more concentrated SFT mixture for the Open-o3 + MCoT variant because the Open-o3 base model has already been exposed to the broader STGR-style training recipe, including SFT and RL on this same corpus family, in its original training pipeline [16]. Accordingly, for the MCoT variant we bias SFT more strongly toward motion-intensive grounded examples, with the goal of adapting the already grounded base model toward more explicit trajectory reasoning rather than repeating the full broad-coverage mixture used to establish its general grounded reasoning ability.

Key difference in training emphasis. The main distinction between the two models lies in the SFT stage. Motion-o is trained with a larger and more mixed SFT corpus that retains substantial general video QA supervision alongside grounded spatio-temporal data, whereas Open-o3 + MCoT uses a smaller but much more concentrated motion-oriented SFT set. Motion-o prioritizes a balance between broad video reasoning competence and explicit motion-aware grounding, while Open-o3 + MCoT prioritizes targeted adaptation of an already grounded base model toward trajectory-explicit reasoning. This also explains why Open-o3 + MCoT achieves higher overall results: the Open-o3

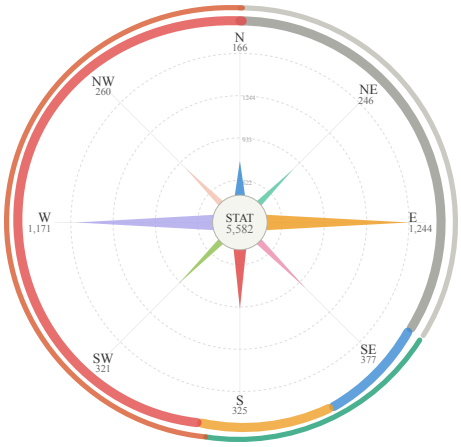


Figure S5: Motion data compass showing the distribution of observations across direction, speed, and scale. Directional wedges radiate from the center using square-root scaling: **N** (166), **NE** (246), **E** (1,244), **SE** (377), **S** (325), **SW** (321), **W** (1,171), and **NW** (260); the central circle is stationary (5,582). The inner ring encodes speed: **stationary** (5,582), **slow** (1,396), **moderate** (1,266), **fast** (1,448). The outer ring encodes scale: **stable** (5,751), **approaching** (2,322), **receding** (1,619). Arc lengths are proportional to each category's share.

base model has already been trained on the STGR corpus [16], giving it a strong prior over spatio-temporal grounding tasks. Our SFT stage therefore does not need to re-establish that general competence, and instead focuses narrowly on introducing `<motion/>` tag reasoning. The subsequent RL stage then builds directly on this foundation, using trajectory-consistency and visual grounding rewards to sharpen the model's ability to produce track-consistent, explicitly grounded motion claims. The performance advantage of Open-o3 + MCoT thus reflects this compounding of a well-initialized base with targeted motion supervision, rather than any fundamental architectural difference.

A.2 Implementation Details

The SFT stage runs for 3 epochs on a single NVIDIA H200 GPU (140GB). The RL stage is conducted using 2 NVIDIA H200 GPUs (140GB).

B Motion Tag Accuracy on the Augmented Dataset

Table S6 evaluates whether the generated `<motion/>` tags are not only syntactically well-formed, but also aligned with the trajectory-derived motion labels in our augmented dataset. We compare predicted tags against ground-truth descriptors computed from dense bounding-box tracks and report accuracy for three attributes: direction (*dir*), speed, and scale. For each attribute, we include both *exact match*, where the predicted bin must equal the ground-truth bin, and *adjacent match*, where predictions within one ordinal step are counted as correct. For direction, adjacent match corresponds to a tolerance of $\pm 45°$; for speed and scale, it corresponds to ± 1 rank. This relaxed metric is important because small localization errors or frame-level jitter can shift a descriptor into a neighboring bin even when the predicted motion is qualitatively reasonable.

The results show that Motion-o learns motion tags that are meaningfully aligned with dense trajectory evidence. In particular, adjacent-match accuracy is consistently higher than exact-match accuracy, indicating that many errors are near misses rather than completely incorrect motion interpretations. This is expected because the motion descriptors are derived from discretized bounding-box trajectories: small errors in object localization, centroid estimation, or box-area change can alter the final direction, speed, or scale bin.

An important observation is that the improvement from the Qwen2.5-based variant to the Qwen3-based variant is modest. We interpret this as evidence that motion-tag accuracy is not limited only by the model's general language or reasoning capacity. The oracle-track control in Appendix H.1 provides more direct evidence that localization quality is an important bottleneck within the current

Table S6: **<motion/> tag accuracy** on the augmented trajectory-grounding dataset. We report exact-match accuracy (%) for each motion attribute. "Avg" denotes the macro-average over three attributes. The r_{ground} column indicates whether the visual grounding reward is enabled (👁) or disabled (🚫). Rows marked *SFT only* use the cold-start supervised stage without reinforcement learning.

Model	r_{ground}	Direction	Speed	Scale	Avg
SFT cold-start only (no RL)					
Motion-o (Qwen2.5-VL-7B)		51.3	68.2	71.4	63.6
Motion-o (Qwen3-VL-8B)		52.8	69.7	72.6	65.0
Open-o3 + MCoT		68.1	79.4	80.2	75.9
SFT + RL (full pipeline)					
Motion-o (Qwen2.5-VL-7B)	🚫	72.1	81.4	82.7	78.7
Motion-o (Qwen2.5-VL-7B)	👁	78.9	88.3	88.4	85.2
Motion-o (Qwen3-VL-8B)	🚫	73.5	83.1	84.2	80.3
Motion-o (Qwen3-VL-8B)	👁	79.6	89.1	89.3	86.0
Open-o3 + MCoT	🚫	74.8	84.2	85.6	81.5
Open-o3 + MCoT	👁	**80.4**	**90.7**	**90.1**	**87.1**

label space. The model may understand *what* object is relevant and *where* it approximately appears, but the derived motion tag depends on precise changes in the predicted box center and area over time. If the bounding boxes are spatially noisy, temporally inconsistent, or only coarsely aligned with the object, then the resulting direction, speed, and scale descriptors can be incorrect even when the model's semantic understanding of the event is broadly correct.

In the current pipeline, bounding boxes are generated by the VLM itself rather than by a dedicated detector or tracker. As a result, the system is constrained by the localization accuracy of the base model. This limitation is especially visible for attributes such as direction and scale, where small spatial errors can flip the discretized label. The ablation without the visual grounding reward further supports this interpretation. Removing r_{ground} weakens the alignment between predicted motion tags and trajectory-derived labels, suggesting that the reward helps encourage motion claims that are more dependent on visual temporal evidence. However, the remaining errors indicate that reward-based verification cannot fully compensate for inaccurate localization. The quality of the motion reasoning trace therefore depends on both components: the model must learn to express motion explicitly, and the underlying visual grounding must provide sufficiently accurate object tracks for those expressions to be reliable.

C Zero-Shot <motion/> Prompting

Table S7 evaluates whether the <motion/> tag schema alone, without any fine-tuning, is sufficient to elicit accurate motion reasoning. We prompt GPT-4o and Qwen2.5-VL-7B zero-shot with and without the full tag specification. While structured prompting yields modest gains over the unstructured baseline, the results fall substantially short of Motion-o's fine-tuned performance (Table S6), confirming that format compliance alone does not substitute for trajectory-grounding supervision and the visual grounding reward.

D Token Overhead of MCoT Responses

MCoT makes motion reasoning explicit by adding structured <motion/> tags to the model's spatio-temporal evidence chain. This increases the length of generated responses compared to baselines that only produce an answer or a standard evidence trace. The results show that MCoT increases response length relative to the base VLM, as expected, because the model emits explicit spatio-temporal evidence and structured <motion/> descriptors rather than only a short final answer. However, this increase should be interpreted as the cost of making reasoning auditable. Motion-o processes more visual evidence and exposes the motion relation that would otherwise remain implicit.

Table S7: **Zero-shot `<motion/>` prompting** evaluated on motion tag accuracy (%) (using the MCoT augmented subset of data) and V-STAR (10% of the dataset). "w/ prompt" denotes zero-shot inference with the full `<motion/>` schema; "w/o prompt" is the standard baseline. No fine-tuning is applied.

Model	Prompt	Motion Tag Acc (%)				V-STAR		
		Dir	Speed	Scale	Avg	Temp tIoU	R@0.3	Spatial mIoU
GPT-4o	w/o `<motion/>`	0.0	0.0	0.0	0.0	23.6	28.0	10.3
GPT-4o	w/ `<motion/>`	**31.9**	**29.8**	**30.5**	**30.7**	**25.4**	**29.0**	**12.5**
Qwen2.5-VL-7B	w/o `<motion/>`	0.0	0.0	0.0	0.0	14.6	18.2	9.3
Qwen2.5-VL-7B	w/ `<motion/>`	**24.1**	**22.3**	**23.8**	**23.4**	**17.1**	**21.4**	**12.7**

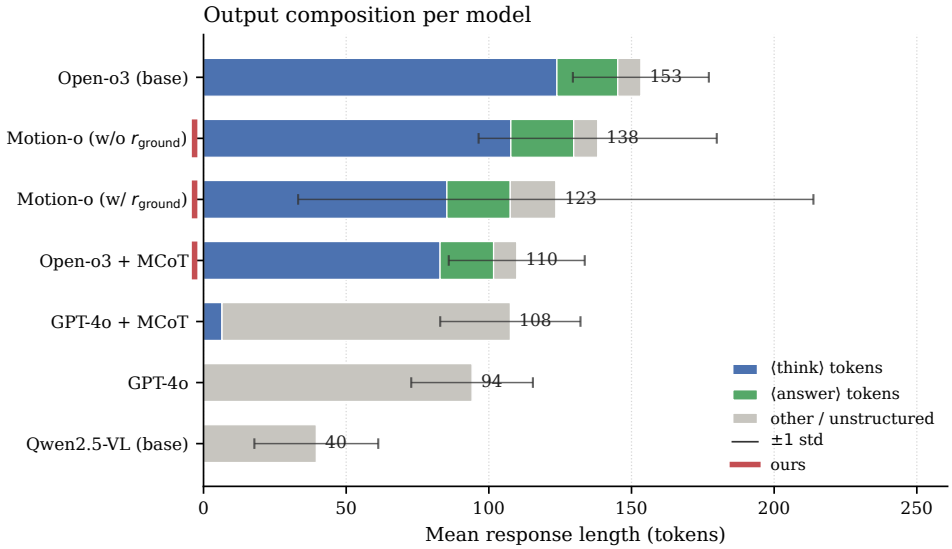


Figure S6: **Generated-token overhead across model variants.** We report the mean response length for each model (ran over 100 samples), decomposed into `<think>` tokens, `<answer>` tokens, and other unstructured tokens, with error bars showing ± 1 standard deviation. Although Motion-o introduces explicit structured tokens for motion grounding, both Motion-o variants and Open-o3 + MCoT remain shorter than Open-o3 base — Motion-o (w/ r_{ground}) uses **19% fewer tokens** (123 vs. 153) and Open-o3 + MCoT uses **28% fewer** (110 vs. 153) — while providing rewardable trajectory evidence, indicating that MCoT supervision densifies rather than lengthens the reasoning trace.

Importantly, the trained MCoT variants remain token-efficient relative to strong grounded-reasoning baselines. Our MCoT-trained models produce responses in the range of 110-138 tokens on average, compared to 153 tokens for the Open-o3 base model. Thus, although Motion-o uses more tokens than a minimal answer-only baseline, it uses approximately 19% fewer tokens than Open-o3 while providing explicit motion-grounded reasoning. This suggests that MCoT does not simply improve performance by making the model generate longer explanations; instead, trajectory supervision encourages more compact and structured reasoning traces.

We also observe that the visual grounding reward improves token efficiency. The variant without r_{ground} produces substantially longer outputs, indicating that removing the grounding signal can lead to less concise reasoning. In contrast, the full Motion-o model learns to express the relevant trajectory information more directly through structured motion tags.

E Extended Benchmark Results

Table 2 reports results on three additional benchmarks: **TVGBench** [26] for temporal video grounding, **MVBench** [12] for multi-modal video understanding, and **MotionBench** [11] (dev split) for

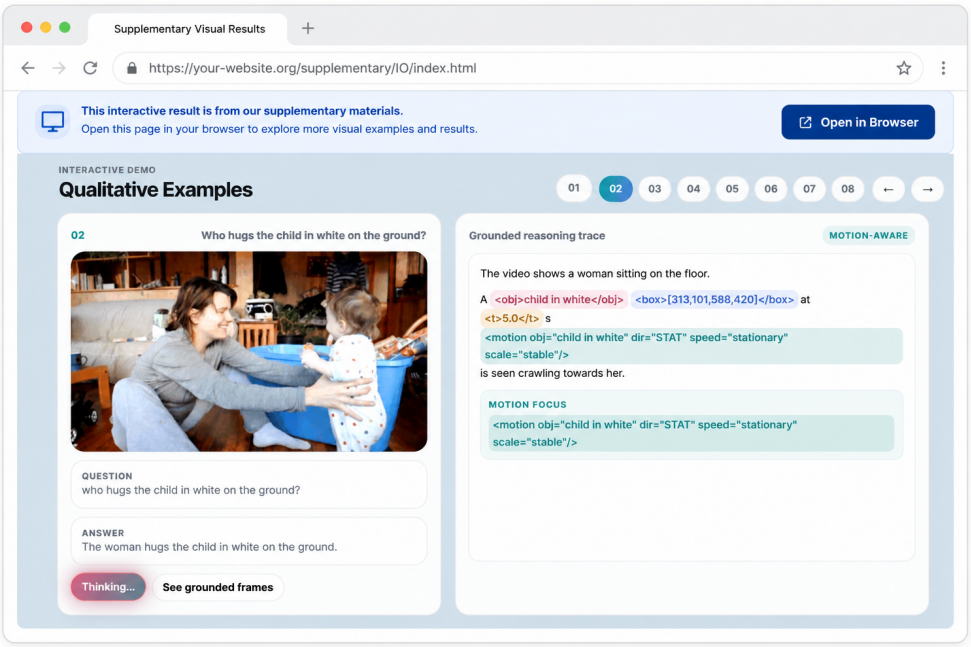


Figure S7: **Additional interactive qualitative results.** We provide an interactive HTML in the supplementary materials that allows you to browse additional visual examples, grounded reasoning traces, motion-aware annotations, and grounded frames. Please open `supplementary/index.html` in a web browser to view the full set of qualitative results beyond the examples shown in the paper.

motion-specific question answering. These complement the V-STAR, VideoMME, and WorldSense results in the main paper.

Quantitative Analysis The benchmark results in Table 2 show a clear and consistent pattern: making motion explicit improves performance not only on motion-focused benchmarks, but also on broader video reasoning tasks that implicitly depend on dynamic understanding. Starting from the Qwen2.5-VL-7B baseline, Motion-o improves TVGBench from 16.3 to 37.6 (+21.3) and MotionBench from 35.0 to 60.0 (+25.0). When added on top of the Open-o3 base model, the Open-o3 + MCoT (Motion Chain-of-Thought) variant further raises TVGBench from 20.8 to 39.6 (+18.8), MVBench from 64.4 to 69.2 (+4.8), and MotionBench from 45.0 to 63.0 (+18.0). Notably, the best variant also surpasses the strongest open-source MVBench baseline in our table (69.2 vs. 67.9) and exceeds GPT-4o result on MotionBench by a wide margin (63.0 vs. 33.0). These gains suggest that explicit trajectory reasoning is improving a capability that many modern video benchmarks increasingly require. This pattern is important because the three benchmarks probe different surfaces of the same underlying problem. TVGBench emphasizes when evidence occurs and whether the model can temporally align its reasoning to the relevant segment. MotionBench is more directly motion-centric, explicitly testing categories such as action order, camera motion, location-related motion, and motion recognition. MVBench is broader and more heterogeneous, but many of its subsets also depend strongly on dynamics rather than static appearance alone, including action sequence, action prediction, and object interaction. Benchmarks are increasingly converging toward testing a model's ability to reason about temporal change, transitions, and object evolution, rather than merely recognizing static content.

The key point we want to emphasize here is that benchmark construction has already started moving in a motion-centric direction, but most models have not. Many recent methods are designed as generic video reasoners: they may improve temporal grounding, spatial localization, or chain-of-thought style inference, but they still do not explicitly model the trajectory that connects observations. Existing evidence-based video reasoning frameworks have become increasingly strong at grounding where an object is and when it appears, yet they typically leave how it moved between those observations implicit. This forces the model to interpolate dynamics internally, often relying

on priors, textual smoothness, or weakly grounded heuristics rather than an explicit motion representation. Motion-o directly targets this gap by introducing with the structured `<motion/>` tag that summarizes direction, speed, and scale change between grounded observations, turning trajectory reasoning into an explicit and rewardable intermediate step. The gains do not come from architectural novelty, but from changing what the model is required to represent. Motion-o changes the supervision target and the reward structure so that motion is no longer an implicit byproduct of spatio-temporal grounding, but an explicit reasoning primitive. Prior models may include motion-sensitive metrics or temporal supervision, but the models themselves do not articulate motion patterns in their reasoning traces. In Motion-o, the trajectory is part of the reasoning trace itself, and the RL stage further encourages those motion claims to be both track-consistent and visually grounded under motion masking.

Qualitative Analysis Figure S9 highlights a successful case where Motion-o captures a subtle but important temporal pattern: the baby moves in one circular direction and later reverses direction. In the generated reasoning trace, the motion grounding reflects this change explicitly, first marking the baby with westward motion and later with eastward motion. This example is qualitatively important because the key evidence is not merely the baby's presence or spatial location, but the change in trajectory over time. The model does not rely only on static snapshots; instead, it uses motion-grounded evidence across timestamps to support the answer. This illustrates the central motivation of Motion-o: making trajectory evolution explicit allows the model to reason about dynamic events that would otherwise remain implicit in standard spatio-temporal grounding pipelines. In this case, the model correctly tracks the temporal progression of the action and exposes the reversal through the structured motion tag, showing that the proposed motion-centric reasoning pathway can capture directional transitions that are essential for true video understanding.

Figure S10 illustrates an important limitation of the current approach. Although the model arrives at the correct motion understanding the spatial localization is weak, as the predicted boxes do not tightly align with the referenced objects. This behavior is consistent with the design of the system. In the current pipeline, bounding boxes are produced by the underlying VLM backbone (here, Qwen), which is not a native object detector and is therefore not optimized for precise localization. As a result, Motion-o can still leverage coarse grounding and motion-aware reasoning to reach the correct answer, but the visual grounding itself may be spatially noisy. This suggests a natural direction for future improvement. One promising extension would be to integrate a stronger spatial localization component, such as a dedicated detector or grounding module, so that trajectory reasoning can operate on more accurate object tracks.

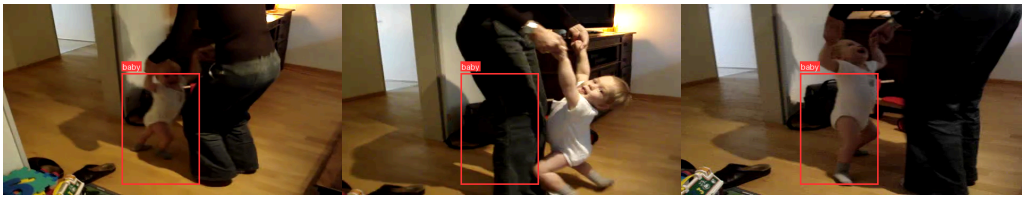
Stationary Motion as Informative Evidence. A natural concern with the `<motion/>` schema is that the `STAT/stationary/stable` configuration might behave as a trivial default; a label the model emits whenever it cannot detect movement, contributing little to downstream reasoning. Figure S8 shows that this is not the case in practice. The question asks what is visible in the background of the video *when the miniature bottle is shown empty*, which requires the model to verify that a candidate object is present *across* the relevant evidence window, not merely at a single keyframe. The model first grounds a hand near the bottle at $t{=}0\,\mathrm{s}$ and then grounds the seashells in the background at $t{=}20\,\mathrm{s}$, and crucially attaches a stationary motion tag to the seashells: `<motion obj="seashells" dir="STAT" speed="stationary" scale="stable"/>`. In the model's own reasoning trace, this tag is explicitly used to justify the claim that the seashells *remain in the frame*, which directly supports the final answer (**C. Shells**) over the distractors (water, cork, star).

This example highlights an asymmetric but important property of MCoT. Directional motion tags (e.g., `dir="W"` with `speed="moderate"`) make the model commit to a specific trajectory pattern between observations, and the trajectory and visual grounding rewards penalize disagreement with the underlying tracks. Stationary tags, by contrast, make the model commit to a different but equally falsifiable claim: that the object's position, velocity, and apparent scale are all approximately invariant across the evidence window. Because our trajectory reward is computed against ground-truth bins derived from centroid displacement and box-area change (Sec. 3.3), an incorrectly emitted `STAT/stationary/stable` triple is penalized just like an incorrect directional triple. Conversely, when the underlying object truly is static, the dual-chain visual grounding reward does not require the prediction to flip under motion masking, since freezing frames does not alter a genuinely stationary scene; the prediction is treated as fully grounded by construction. The stationary configuration therefore functions as a positive, rewardable assertion of temporal persistence rather than as an ab-



Question: Which of the following is visible in the background of the video when the miniature bottle is shown empty?
Reasoning: <think>The video shows a miniature glass bottle being held by a hand. In the background, various objects are visible. At the beginning of the segment, a <obj>hand</obj><box>[234,156,420,287]</box>at<t>0</t>s is holding the bottle, and several seashells<obj>seashells</obj><box>[0,209,420,364]</box> can be seen scattered around. As the hand tilts the bottle, more seashells become visible in the background. Later, the hand places the bottle down, and the <obj>seashells</obj><box>[0,209,420,364]</box>at<t>20</t>s<motion obj="seashells" dir="STAT" speed="stationary" scale="stable"/> remain in the frame, confirming their presence.</think>
Answer: A. Water. B. Cork. **C. Shells.** ✅ D. Star.

Figure S8: **Stationary motion as informative evidence.** Even when an object does not move, the explicit <motion/> tag carries non-trivial reasoning value. Here, the model grounds the seashells in the background and emits <motion obj="seashells" dir="STAT" speed="stationary" scale="stable"/>, certifying that the seashells *remain in the frame* across the evidence window rather than appearing only briefly. This persistence claim is precisely what the question requires ("visible in the background ... when the miniature bottle is shown empty"), and it allows the model to confidently select **C. Shells** over the distractors. The example illustrates that dir="STAT" is not a degenerate or empty label: it is an explicit, rewardable assertion of temporal persistence that complements the directional motion cases shown in the main paper.



Question: Who does the adult hold the hand of?

Reasoning: The video shows an adult helping a baby learn to walk. At the beginning of the clip, the <obj>baby</obj><box>[223,130,370,340]</box>at<t>1.5</t>s is holding onto the adult's hands for support as they take their first steps. As the baby continues to practice walking, the adult holds the baby's hands to help them maintain balance. For instance, the <obj>baby</obj><box>[223,130,370,340]</box>at<t>3.0</t>s<motion obj="baby" dir="W" speed="slow" scale="stable"/> is still holding the adult's hands while learning to walk. This interaction continues throughout the segment, as seen again when the <obj>baby</obj><box>[223,130,370,340]</box>at<t>12.0</t>s<motion obj="baby" dir="E" speed="slow" scale="stable"/> is holding the adult's hands. Therefore, the adult is holding the hand of the baby.

Answer: The adult holds the hand of a baby.

Figure S9: Qualitative example showing successful motion-centric reasoning: the model captures the baby's change in movement direction over time through explicit motion grounding.

sence of information, which is why it can carry the weight of the answer in cases like Figure S8. This complements the directional examples shown in the main paper (Figure 4) and in Figure S9, and supports our broader claim that explicit motion descriptors, whether dynamic or static, convert implicit trajectory assumptions into verifiable evidence.

F Prompt Templates

To support full reproducibility, this section documents the exact message structures used during supervised fine-tuning (SFT), GSPO training, and benchmark evaluation. All prompts follow the TRL [22] standard chat format consumed by Qwen2.5-VL [3] and Qwen3-VL [2]. Placeholders of the form <...> are filled in per-example at runtime, and the structured tag *<motion/>* denotes the trajectory descriptor introduced in Sec. 3.2.



Question: What is beneath the adult wearing red pants?

Reasoning: The video shows an adult water skiing on a body of water. The adult is wearing red pants and a life vest while being pulled by a rope attached to a boat. Throughout the segment, the `<obj>water skier</obj><box>[38,101,181,224]</box>at<t>1.7</t>s` is seen gliding over the surface of the water. The water skier's feet are firmly planted on the `<obj>skis</obj><box>[38,164,181,224]</box>at<t>1.7</t>s<motion obj="skis" dir="W" speed="slow" scale="stable"/>`, which are positioned directly above the water. This visual evidence confirms that the water skier is standing on the skis, which are on the water.

Answer: The adult is standing on a pair of skis, which are on the water.

Figure S10: Qualitative limitation example: the model reaches the correct answer, but the spatial grounding is imprecise, highlighting the need for stronger localization modules.

F.1 Frame Sampling

For every training and evaluation example involving a video, we uniformly sample N frames from the video segment associated with the question and pass them in temporal order, denoted *<frame_1>*, ..., *<frame_n>*. We use N=32 frames during training and N=64 at evaluation unless otherwise noted. The prompt structure is independent of N. For the dual-chain visual grounding reward (Sec. 3.3), the *motion-masked* variant *<masked_video_frames>* replaces intermediate frames with frozen copies of keyframes to remove temporal motion cues while preserving appearance.

F.2 Training Prompts (Task-Conditioned)

Training samples are built with a task-conditioned *system* message followed by a multimodal *user* message. The system message specifies the output grammar that our reward parsers (format, trajectory, and visual grounding) consume; we therefore list each variant verbatim. Tasks correspond to the SFT/RL mixture described in Appendix A.

Temporal-Spatial Free-Form QA

This is the primary trajectory-grounding task: the model must ground each observation with `<obj>`, `<box>`, `<t>` tags and emit a self-closing *<motion/>* descriptor after the last mention of any object that appears at ≥ 2 timestamps.

SFT/GSPO Training Message — Temporal-Spatial Free-Form QA

```
messages = [
    {
        "role": "system",
        "content": [
            {
                "type": "text",
                "text": (
                    "A conversation between user and assistant. The user provides a
    video "
```

```
                    "and asks a question, and the assistant solves it by reasoning over "
                    "spatio-temporal evidence.\n"
                    "The reasoning process and answer are enclosed within "
                    "<think> </think> and <answer> </answer> tags respectively.\n"
                    "Inside <think>, every grounded observation must be written as "
                    "<obj>name</obj><box>[x1,y1,x2,y2]</box>at<t>seconds</t>s.\n"
                    "After the last grounded mention of an object that appears at >=2 "
                    "timestamps, emit a self-closing motion tag:\n"
                    "  <motion obj=\"name\" dir=\"D\" speed=\"S\" scale=\"C\"/>\n"
                    "where:\n"
                    "  dir   in {N, NE, E, SE, S, SW, W, NW, STAT}\n"
                    "  speed in {stationary, slow, moderate, fast}\n"
                    "  scale in {approaching, stable, receding}\n"
                    "The <answer> block contains plain text only "
                    "(no <obj>, <box>, or <t> tags)."
                ),
            },
        ],
    },
    {
        "role": "user",
        "content": [
            {"type": "video", "video": <video_frames>},
            {"type": "text",  "text": <question_text>},
        ],
    },
]
```

General Video QA (MCQ)

Multiple-choice general video QA reuses the *<motion/>* schema inside `<think>`, but constrains `<answer>` to a single option letter.

SFT/GSPO Training Message — General Video QA (MCQ)

```
messages = [
    {
        "role": "system",
        "content": [
            {
                "type": "text",
                "text": (
                    "A conversation between user and assistant. The user provides a
    video, "
                    "a question, and a list of options. The assistant reasons inside "
                    "<think>...</think>, including <motion/> tags whenever motion is "
                    "relevant, and outputs only the option letter inside "
                    "<answer>...</answer>."
                ),
            },
        ],
    },
    {
        "role": "user",
        "content": [
            {"type": "video", "video": <video_frames>},
            {"type": "text",
             "text": "Question: <question_text>\nOptions:\n<options_text>"},
        ],
    },
]
```

Temporal QA

Temporal grounding samples ask the model to localize *when* an event occurs and constrain `<answer>` to a fixed time-range format. No *<motion/>* tag is required.

SFT/GSPO Training Message — Temporal QA

```
messages = [
    {
        "role": "system",
        "content": [
            {
                "type": "text",
                "text": (
                    "A conversation between user and assistant. The user provides a
    video "
                    "and asks when an event occurs.\n"
                    "The assistant first thinks inside <think>...</think> and then "
                    "provides the final answer inside <answer>...</answer>.\n"
                    "The <answer> block must follow the format:\n"
                    "  From <t>t1</t>s to <t>t2</t>s"
                ),
            },
        ],
    },
    {
        "role": "user",
        "content": [
            {"type": "video", "video": <video_frames>},
            {"type": "text",  "text": <question_text>},
        ],
    },
]
```

Visual QA (Image)

Single-image grounding samples (used to maintain spatial-grounding competence) require `<obj>/<box>` grounding inside `<think>`, but no temporal or motion tags.

SFT/GSPO Training Message — Visual QA (Image)

```
messages = [
    {
        "role": "system",
        "content": [
            {
                "type": "text",
                "text": (
                    "A conversation between user and assistant. The user asks a question
    "
                    "about an image, and the assistant solves it.\n"
                    "The assistant first thinks inside <think>...</think> and then "
                    "provides the final answer inside <answer>...</answer>.\n"
                    "Inside <think>, every visual claim must be grounded as "
                    "<obj>name</obj><box>[x1,y1,x2,y2]</box>."
                ),
            },
        ],
    },
    {
        "role": "user",
        "content": [
            {"type": "image", "image": <image>},
            {"type": "text",  "text": <question_text>},
```

```
        ],
    },
]
```

F.3 Inference Prompt (Unified)

At inference time, our vLLM wrapper uses a single unified system prompt (`MOTION_SYSTEM_PROMPT`) for all trajectory-grounded video reasoning, regardless of whether the downstream task is free-form or multiple-choice. This matches the temporal-spatial free-form QA training distribution most closely and is also used by the dual-chain motion masking pass: both the original and the motion-masked rollouts are conditioned on the same system message, so any change in predicted *<motion/>* attributes is attributable to the temporal evidence rather than to prompt asymmetry.

Inference Message — Unified Trajectory-Grounded Prompt

```
messages = [
    {
        "role": "system",
        "content": [
            {
                "type": "text",
                "text": (
                    "A conversation between user and assistant. The user provides a
    video "
                    "and asks a question, and the assistant solves it.\n"
                    "The assistant first thinks about the reasoning process in the mind "
                    "and then provides the user with the answer.\n"
                    "Inside <think>...</think>, the assistant grounds each observation
    as "
                    "<obj>name</obj><box>[x1,y1,x2,y2]</box>at<t>t</t>s.\n"
                    "After grounding an object across multiple timestamps, the assistant
    "
                    "emits <motion obj=\"name\" dir=\"D\" speed=\"S\" scale=\"C\"/> "
                    "summarizing the trajectory, where:\n"
                    "  dir   in {N, NE, E, SE, S, SW, W, NW, STAT}\n"
                    "  speed in {stationary, slow, moderate, fast}\n"
                    "  scale in {approaching, stable, receding}\n"
                    "The reasoning process and answer are enclosed within "
                    "<think> </think> and <answer> </answer> tags respectively.\n"
                    "The answer part only requires a text response; tags like "
                    "<obj>, <box>, <t> are not needed."
                ),
            },
        ],
    },
    {
        "role": "user",
        "content": [
            {"type": "video", "video": <video_frames>},
            {"type": "text",  "text": <question_text>},
        ],
    },
]
```

F.4 Benchmark Evaluation Prompts

For benchmark evaluation, we follow each benchmark's official prompting convention to ensure direct comparability with prior work. The **V-STAR** benchmark explicitly evaluates spatio-temporal grounding and reuses the unified inference prompt above. The remaining benchmarks (MVBench, MotionBench, TVBench, VideoMME, WorldSense) constrain the model to emit only a single option

letter; the model is *not* required to produce <obj>/<box>/<t> or <motion/> tags at the output level, though it may still use them internally inside <think>.

Evaluation Message — MVBench / MotionBench / TVBench (Letter-Only MCQ)

```
messages = [
    {
        "role": "system",
        "content": [
            {
                "type": "text",
                "text": (
                    "Carefully watch the video and pay attention to every detail. "
                    "Based on your observations, select the option that best answers "
                    "the question. Answer with only the letter of your choice."
                ),
            },
        ],
    },
    {
        "role": "user",
        "content": [
            {"type": "video", "video": <video_frames>},
            {"type": "text",
             "text": "Question: <question_text>\nOptions:\n<options_text>"},
        ],
    },
]
```

F.5 Summary of Prompt Components

Component	Training / Inference (Trajectory-Grounded)	Benchmark Evaluation (Letter-Only MCQ)
Frames	N uniformly sampled, temporal order	N uniformly sampled, temporal order
System message	Task-conditioned grammar (Sec. F.2) or unified inference prompt (Sec. F.3)	"Watch carefully … answer with only the letter"
<think> requirements	<obj>/<box>/<t> grounding + <motion/> tags	Free (not enforced at parse time)
<answer> content	Plain text, option letter, or From <t>t_1</t>s to <t>t_2</t>s (task-dependent)	Single option letter
Output target	Free-form generation with structured tags	Single letter $\in$ {A, B, C, D}
Scoring	$r_{\text{acc}} + r_{\text{thk}} + r_{\text{fmt}}$ (Sec. 3.3)	Exact-match top-1 accuracy

Table S8: Differences between trajectory-grounded prompts (used for SFT, GSPO rollouts, and V-STAR/free-form inference) and letter-only benchmark-evaluation prompts. The trajectory-grounded prompts define the output grammar that Motion-o's reward parsers consume; the benchmark-evaluation prompts follow each benchmark's official protocol to ensure faithful comparison.

G Broader Impacts

This work aims to get one step closer to improving the trustworthiness of VLMs by making motion-based reasoning more explicit, interpretable, and verifiable. By requiring models to expose trajectory-level evidence through structured <motion/> tags, Motion-o provides a mechanism for checking whether dynamic claims are supported by observed changes in the video rather than by static appearance cues or language priors. This direction may be especially valuable for domains where video understanding must be reliable, such as assistive technologies, robotics, healthcare monitoring, transportation, and other safety-critical settings.

A positive impact of this work is that explicit motion reasoning can make model outputs easier to audit. Instead of only producing a final answer, the model provides intermediate motion descriptors that can be compared against object tracks or perturbed inputs. This creates opportunities for verification and failure detection before a model response is trusted in downstream use. More broadly, methods that encourage models to verify their responses against visual evidence may support safer deployment of VLMs in settings where incorrect temporal or causal interpretations could have serious consequences.

At the same time, improved video understanding can also introduce risks. More reliable trajectory reasoning could be misused in privacy-sensitive surveillance, tracking, or behavioral analysis applications. Motion-o does not introduce new surveillance data or deploy models in such settings, but the underlying capability of more accurate motion interpretation should be handled carefully. We therefore view this work as primarily a step toward more transparent and accountable video reasoning, and we encourage future deployments to include appropriate privacy protections and task-specific validation before use in high-stakes environments.

H Oracle Tracks, Masked-Chain Diagnostics, and Reward Variants

H.1 Oracle-Box Control and Representation Limits

Table S9 replaces the model's predicted boxes with ground-truth tracks at inference. Tag accuracy improves by 11.6 points and fine-motion subset accuracy by 22.1. Reliable localization is therefore an important constraint within the current label space. However, the remaining 15.5-point fine-motion gap prevents a claim that localization explains every failure. This experiment does not supply a breakdown by schema expressibility or show that curved, multi-segment, occluded, interacting, or camera-induced motion can be represented by one global direction/speed/scale tag. It complements the predicted-box arithmetic control: clean tracks make the labels easier to recover, while emitted tags outperform arithmetic on noisy predicted tracks.

Table S9: Oracle-box control (accuracy, %). Fine/compound-motion accuracy is reported for the aggregate probe.

Box source	Tag accuracy (macro avg.)	V-STAR fine-motion subset
Predicted boxes	85.2	62.4
Oracle ground-truth tracks	96.8	84.5

H.2 What the Motion-Masked Chain Produces

The motion-masked input freezes only the spans associated with motion tags, using repeated keyframe copies rather than freezing the entire video. The original and masked inputs share the same unified prompt; the perturbation changes the visual input, not the instructions. Table S10 reports the logged distribution over masked completions from the RL evaluation set. Stationary tags dominate, which is consistent with frozen motion evidence, but this is a diagnostic rather than direct accuracy supervision for Chain B. The changed non-stationary category is not independently labeled as correct or incorrect.

Table S10: Masked-completion output distribution logged during RL evaluation. Percentages sum to 100.0.

Masked completion	Share (%)
Transitioned to `dir="STAT"`	81.4
Changed attribute, non-STAT	7.2
Unchanged tag	4.1
Absent or dropped object	7.3

The original reward credits absent objects as grounded, so dropping an object can receive credit without demonstrating a correct counterfactual. The observed 7.3% absence rate documents its incidence; it does not prove that reward exploitation is impossible. Assigning zero credit to absent objects would remove this particular shortcut and is a straightforward safeguard for future implementations. That safeguard was not used in the experiments reported here. Original-chain trajectory correctness remains supervised by r_{traj}.

H.3 Preliminary Distance-Based Grounding Reward

The perturbation principle can be formulated with a continuous quantity. For example, if D denotes centroid distance traveled, a possible score is

$$r_D = \text{clip}\left(\frac{D_{\text{real}} - D_{\text{frozen}}}{D_{\text{real}} + \epsilon}, 0, 1\right).$$

This illustrates rewarding motion that disappears when temporal evidence is removed; it is not a claim that all continuous motion targets have been validated. The preliminary run in Table S11 changes the formulation of r_{ground} alone, retains the discrete tag and active r_{traj} in every row, and

uses 4K training samples for 400 LoRA steps. V-STAR uses 200 of 2,094 examples with Qwen2.5-72B as judge; VideoMME uses the full evaluation set. These absolute values are not comparable to the full-data main results.

Table S11: Small-scale grounding-reward pilot. All rows retain discrete MCoT tags and trajectory supervision; "no reward" omits only the perturbation reward.

Grounding reward	V-STAR mAM	V-STAR mLGM	VideoMME
No perturbation reward	31.2	44.2	62.2
Continuous, distance-based	32.5	45.8	63.3
Discrete, categorical default	33.8	47.6	64.6

Both reward formulations improve the reported scores over omitting the perturbation term. At this scale, differences are preliminary and within the noise of the limited evaluation; they do not establish statistical significance or fully continuous trajectory reasoning. Richer motion outputs, stronger correctness constraints on masked completions, and broader validation remain open directions.

I Limitations of the Motion Interface and Controls

The current descriptor summarizes one object's motion with a global direction, speed bin, and apparent scale change in 2D image space. It cannot faithfully encode a curved or multi-segment path, detailed hand-object interactions, occlusion reasoning, or a decomposition of camera and object motion. Oracle tracks establish a localization effect for the representable labels, not a general solution to these cases. The EgoSchema results are likewise mixed: Motion-o improves over Open-o3 Video but remains below the original Qwen2.5 baseline, while Open-o3 + MCoT exceeds it. Broad egocentric manipulation remains outside the validated scope.

The perturbation reward is a proxy for visual dependence, not a proof of a correct motion claim. It can credit an incorrect attribute change and, in the reported formulation, an absent object. The masked-chain distribution characterizes observed behavior without ruling out these failure modes. The continuous-reward pilot retains discrete tags and uses limited training and judging samples. Finally, the additional controls are aggregate comparisons without repeated-seed uncertainty estimates; the fine/compound-motion oracle probe does not separate representable from schema-inexpressible cases. These limits motivate stronger localization, richer motion representations, and more constrained counterfactual rewards.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction accurately describe Motion-o as a lightweight framework for explicit trajectory-grounded video reasoning through STT reasoning, MCoT, and trajectory-aware training. The method and experiments support these claims through benchmark improvements and ablations.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper discusses limitations such as coarse single-object motion descriptors and weaker spatial localization due to reliance on the underlying VLM for bounding boxes (see Appendices B and I). It also identifies future directions, including multi-object interactions, occlusions, non-linear trajectories, and camera-motion disentanglement.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: The paper does not present formal theoretical results or proofs. The equations define the method, motion descriptors, and reward functions rather than theoretical claims.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper provides the base models, training stages, motion tag schema, data augmentation process, rewards, benchmark suite, prompt templates, and compute details needed to reproduce the main results. These details are further supported by the appendix.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The code is released in the supplementary materials. The augmented trajectories, derived motion descriptors, and dataset splits (built on the existing STGR [16] dataset) will be released to support reproducibility.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The main paper (in 3 and 4) and appendix (in A and F) specify the model, training pipeline, datasets, benchmarks, frame sampling, prompt templates, reward design, and compute resources.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

Justification: The paper reports accuracies across multiple video understanding benchmarks, ablations, and motion-tag accuracy, but does not include error bars or significance tests.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The paper details this information in the appendix in A.2.

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The work uses existing video reasoning datasets and focuses on improving interpretable, trajectory-grounded VLM reasoning. It does not involve new sensitive data collection or human-subject experiments.

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: An explanation for the broader impacts of this work is provided in the Appendix in section G.

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not release a high-risk dataset or model involving sensitive personal data or security-critical deployment. The released assets are trajectory annotations, derived motion descriptors, and splits for existing research datasets.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The paper cites and properly credits the models and datasets for which this work was built on by citing the corresponding technical reports associated with used models/datasets.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The paper documents the newly introduced trajectory-grounding augmentation, including dense trajectories, motion descriptors, and MCoT formatting. The supplementary material also includes code that has been well-documented and clearly detailed.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

Justification: The paper does not conduct crowdsourcing experiments or collect new human-subject data. The annotations are derived from existing spatio-temporal grounding data.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The work does not involve participant recruitment, human-subject experiments, or collection of identifiable participant data.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: The paper states in method section (see Section 3) that Motion-o builds on the Qwen family of VLMs and evaluates VLM-based grounded video reasoning. The experiments (see Section 4) also use additional families of VLMs to compare results across benchmarks.