WHOSE MOTION IS IT? LEARNING MOTION-SOURCE ATTRIBUTION IN VISION-LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Video-language reasoning requires understanding how visual content evolves over time. Yet such changes may arise from camera motion, target-object dynamics, or both, making their physical source ambiguous. We identify a recurring failure mode in current video-language models (VLMs), termed **motion-source misattribution**, where observed changes are attributed to the wrong physical source. Our empirical analysis shows that motion judgments are not reliably bound to the queried target, source information is recoverable but not reliably source-selective, and camera and object evidence interact non-additively when both motions co-occur. These findings reveal a clear gap: encoding motion cues does not guarantee organizing them by physical source. To address this gap, we propose **MoSDeR** (**Mo**tion-**S**ource **De**composition and **R**easoning), which structures temporal evidence into query-conditioned camera and object factors. MoSDeR first binds motion evidence to the queried entity, then routes the two factors into the native reasoning pathway, and finally composes them non-exclusively under concurrent motion. To evaluate motion-source attribution across diverse conditions, we curate a unified target-level benchmark spanning egocentric and allocentric settings, indoor interactions, autonomous-driving scenes, and diverse camera–object motion configurations. Across three VLMs, MoSDeR exceeds the runner-up in Overall accuracy by 3.64% on our benchmark and 6.75% on OmniVCHall on average, indicating consistent attribution gains and broader downstream benefits.[1]

1 INTRODUCTION

Video-language reasoning requires understanding not only what appears in a video, but also how objects, actions, and spatial relations evolve over time. Such temporal reasoning is central to egocentric understanding (Li et al., 2026; Lin et al., 2022), autonomous driving (Wilson et al., 2021), and video question answering (Pătrăucean et al., 2023; Li et al., 2024; Fu et al., 2025). However, visual changes across frames can arise from both scene dynamics and changes in viewpoint, making image-space motion inherently ambiguous. A stationary object may appear to move as the camera moves (see Fig. 1(a)), while a moving object can remain nearly fixed when tracked by the camera (see Fig. 1(b)). Reliable video-language reasoning therefore requires distinguishing viewpoint-induced changes from the physical dynamics of the queried object (Xue et al., 2026).

Despite recent progress in temporal reasoning, separating the physical sources of motion remains challenging. Video architectures strengthen cross-frame modeling to capture evolving appearance and motion patterns (Wang et al., 2024a; Zeng et al., 2025), while trajectory-based representations provide persistent object-centric cues across time (Zheng et al., 2025; 2026). Recent methods incorporate camera dynamics (Lin et al., 2025), geometric structure (Zhou et al., 2026), and region-level or 4D motion cues (Yang et al., 2026; Cho et al., 2026) to support more physically grounded video understanding. As summarized conceptually in Fig. 1(c), these advances improve temporal representation and reasoning, but typically do not explicitly organize motion evidence according to whether it originates from the camera or the queried object. As a result, viewpoint-induced changes and target-object dynamics can still be conflated during reasoning, particularly when both occur simultaneously. This failure is illustrated in Figs. 1(a) and 1(b): strong VLMs can detect temporal changes yet incorrectly attribute them to camera or target-object motion.

We refer to this failure as **motion-source misattribution**: a model observes temporal changes but attributes them to the wrong physical source. To understand why this occurs, we systematically

[1] Our source code is included in the supplementary material for review.

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

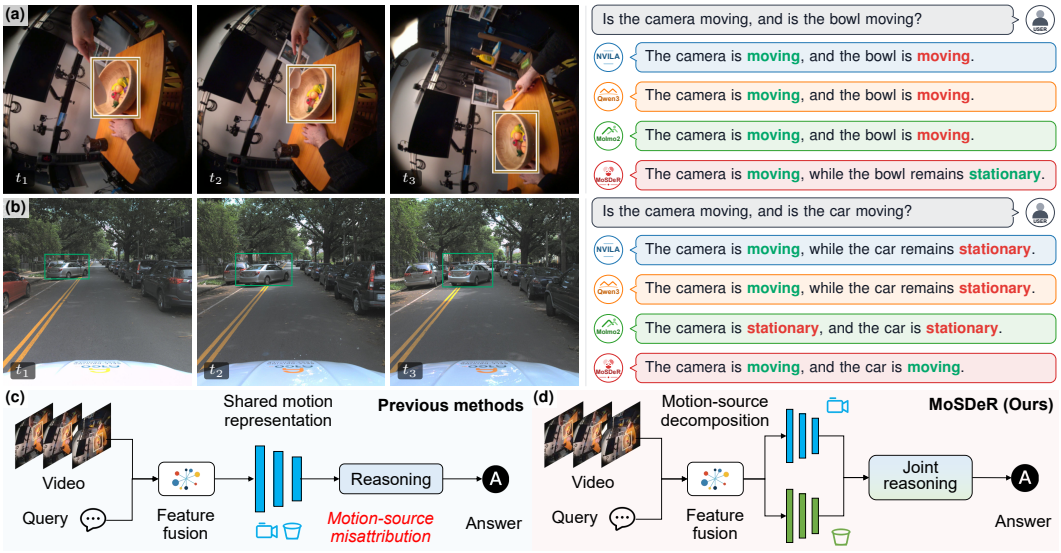


Figure 1: **Motion-source misattribution.** (a)–(b) Existing VLMs can recognize visual motion but misattribute its physical source, confusing camera motion with target-object motion. (c) Prior methods typically reason from shared motion representations in which camera and object cues are not explicitly organized by source. (d) MoSDeR instead decomposes query-conditioned motion evidence into camera and object factors and jointly composes them for reasoning.

diagnose motion-source reasoning across multiple VLM families and uncover three findings. First, models exhibit unreliable *target-conditioned attribution*, indicating that motion evidence is not consistently bound to the queried object (see Figs. 2(a) and 2(b)). Second, source-related information is partially recoverable from intermediate representations, yet it does not reliably exert *source-selective influence* on final decisions (see Fig. 2(c)). Third, camera and object evidence interacts non-additively when both motions co-occur (see Fig. 2(d)). Together, these findings reveal a clear gap: **encoding motion cues does not guarantee organizing them by physical source**.

Motivated by these findings, we propose **MoSDeR** (**Mo**tion-**S**ource **De**composition and **R**easoning), which explicitly organizes temporal evidence by physical source, as illustrated in Fig. 1(d). To address unreliable target-conditioned attribution, MoSDeR first binds motion evidence to the queried entity and derives query-conditioned camera and object factors. To bridge the gap between recoverable source information and final decisions, source-selective adaptation routes the two factors into the native visual and language reasoning pathways. To handle concurrent motion, the factors remain non-exclusive and are jointly composed by the decoder, allowing camera and object dynamics to contribute simultaneously. Together, these designs transform native temporal features into query-aware, source-selective, and composable representations for video-language reasoning.

To evaluate motion-source attribution, we curate a unified target-level benchmark spanning egocentric and allocentric settings, indoor interactions, autonomous-driving scenes, and diverse camera–object motion configurations from four public video datasets: ADT-LiteOffice (Pan et al., 2023), HOT3D (Banerjee et al., 2025), TACO-V1-allocentric (Liu et al., 2024), and AV2 (Wilson et al., 2021). Across NVILA, Qwen3, and Molmo2 (Liu et al., 2025; Bai et al., 2025; Clark et al., 2026), MoSDeR exceeds the runner-up in Overall accuracy by 3.64% on average. On OmniVCHall (Xing et al., 2026), it exceeds the runner-up by 6.75%, confirming downstream transfer.

Our contributions are threefold: (1) we identify and diagnose **motion-source misattribution** in VLMs; (2) we propose **MoSDeR**, which binds motion evidence to the queried target, routes camera and object factors into reasoning, and composes them non-exclusively under concurrent motion; and (3) we curate a target-level benchmark and demonstrate that source-aware motion decomposition improves both source attribution and downstream video-language reasoning.

2 Related Work

Motion Representation and Viewpoint Dynamics. Video understanding requires modeling temporal dynamics together with appearance. Two-stream networks combine RGB appearance with optical flow, while video transformers and masked pretraining learn temporal representations directly

from frame sequences (Simonyan & Zisserman, 2014; Bertasius et al., 2021; Tong et al., 2022). Because image-space motion mixes scene dynamics and viewpoint changes, improved trajectories compensate for camera motion in action descriptors (Wang & Schmid, 2013). Recent work uses camera trajectories as semantic cues (Xue et al., 2026) and object trajectories as persistent representations (Zheng et al., 2025; 2026). For video generation, OrthoMotion disentangles camera and subject controls through orthogonal geometric and semantic attention operators (Meng, 2026). We instead study how these physical sources are attributed and composed for VLM reasoning.

Motion Reasoning in Video-Language Models. Video-language models connect temporal visual representations with language reasoning through multimodal alignment and instruction tuning (Tang et al., 2026). InternVideo2, LLaVA-Video, Video-STaR, and TimeSuite improve temporal representation and video-language alignment (Wang et al., 2024a; Zhang et al., 2025a; Zohar et al., 2025; Zeng et al., 2025), while EgoVLP, LaViLa, and EgoVLPv2 focus on first-person video-language understanding (Lin et al., 2022; Zhao et al., 2023; Pramanick et al., 2023). Meanwhile, MotionBench (Hong et al., 2025), CameraBench (Lin et al., 2025), and DSI-Bench (Zhang et al., 2025b) expose persistent limitations in fine-grained motion, camera behavior, and dynamic spatial reasoning. Geometry, region cues, and fixed-reference supervision support motion reasoning (Zhou et al., 2026; Yang et al., 2026; Cho et al., 2026); MASS injects depth- and tracker-derived 3D signals into the VLM language space (**?**). MoSDeR instead addresses **motion-source misattribution** by deriving query-conditioned camera/object factors from native visual tokens and routing them internally, without an external depth, motion, or trajectory model at inference.

Grounded and Reliable Video-Language Reasoning. VideoHallucer, EventHallusion, and VidHalluc assess unsupported entity, event, temporal, and scene predictions (Wang et al., 2024b; Zhang et al., 2024; Li et al., 2025), while MASH-VLM uses disentangled representations and OmniVCHall introduces decoding-time calibration (Bae et al., 2025; Xing et al., 2026). Know-Show evaluates spatio-temporal grounded reasoning, and its GRAM plug-in augments decoding with attention-selected video tokens and explicit timestamps (Sugandhika et al., 2025). These strategies address representation or decoding-time grounding; MoSDeR instead routes query-conditioned camera and object factors and composes them non-exclusively before answer generation.

3 Empirical Analysis

Motion-Source Reasoning Formulation. Given a video V and a query q about a target entity, a VLM predicts $\hat{a} = f(V, q)$ from the observed temporal visual evidence. We characterize the physical source of this evidence using two non-exclusive binary variables: $y_c \in \{0, 1\}$ for camera motion and $y_o(q) \in \{0, 1\}$ for motion of the queried target relative to the scene. Their combination $s(V, q) = (y_c, y_o(q))$ defines four states: neither, camera-only, object-only, and concurrent motion. Motion-source attribution requires camera evidence to remain invariant across target queries in the same video, while object evidence should follow the corresponding queried entity.

Diagnostic Setup. We study NVILA-Lite-8B (**NVILA**) (Liu et al., 2025), Qwen3-VL-8B-Instruct (**Qwen3**) (Bai et al., 2025), and Molmo2-O-7B (**Molmo2**) (Clark et al., 2026), representing efficient video processing, general multimodal instruction, and open video understanding with grounding, respectively. AV2 provides state-balanced natural coverage of all four camera–target motion configurations, while same-video HOT3D target pairs isolate whether motion judgments follow the queried entity (Wilson et al., 2021; Banerjee et al., 2025). Matched controlled variants isolate camera and target contributions under fixed scene content. Details are provided in **Appendices** B.1 and C.1.

Finding 1: Motion Judgments Are Not Reliably Bound to the Queried Target. We first examine whether motion judgments follow the queried entity. As shown in Fig. 2(a), all three models exhibit clear state-dependent biases: NVILA achieves 73.3% recall on camera-only cases but only 1.7% on object-only cases, while Molmo2 reaches 100.0% on neither-moving cases but performs poorly once motion is present; Qwen3 is more balanced, yet its recall drops to 28.3% under concurrent motion. We then fix the video and switch the queried target. As shown in Fig. 2(b), the camera judgment should remain unchanged while the object judgment should follow the selected target, yet across 21 target pairs, the best model answers both queries fully correctly in only $7/21$ cases. **Appendix** Fig. A2 provides a complementary AV2 example of this target-dependent attribution failure. These results indicate that recognizing motion does not ensure reliable target-conditioned attribution.

Finding 2: Source Information Is Partially Recoverable but Not Reliably Used Selectively. We test whether intermediate representations contain source information that the native model does not use selectively. As shown in Fig. 2(c), full-video linear probes raise four-state accuracy from 28.3%

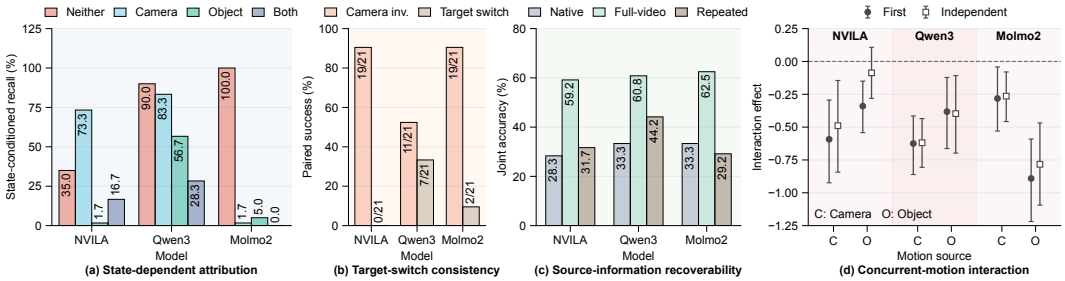


Figure 2: **Diagnosing motion-source misattribution in existing VLMs.** (a) Models exhibit strong state-dependent attribution biases across the four camera–object motion states. (b) Target switching within the same video reveals unreliable target-conditioned attribution, with at most $7/21$ pairs answered fully correctly. (c) Source information is partially recoverable from intermediate representations: full-video probes substantially outperform native predictions, whereas repeated-frame inputs reduce recoverability. (d) Camera and object evidence interacts non-additively under concurrent motion, with the pattern largely reproduced by an independent construction.

to 59.2% for NVILA, 33.3% to 60.8% for Qwen3, and 33.3% to 62.5% for Molmo2. Probe accuracy estimates linearly available source information rather than a competing predictor; our conclusion rests on contrasts with repeated-frame and permutation controls. Repeating one frame reduces probe accuracy to 31.7%, 44.2%, and 29.2%, respectively, showing that temporal variation contributes to recoverability. Yet recoverability does not imply selective decoder use (Hewitt & Liang, 2019; Elazar et al., 2021): under stricter content isolation it varies across model families, while source-direction interventions yield no reliable object-selective effects across families. Further results in **Appendices** B.3 and B.4 expose a gap between recoverable source information and evidence that reliably guides the native model's motion-source decision under these controls.

Finding 3: Concurrent Motion Produces Non-Additive Source Evidence. Finally, we examine how camera and object evidence interacts under concurrent motion. Using matched constructions covering neither, camera-only, object-only, and both-motion states, we measure whether the incremental evidence contributed by one source changes when the other is present. As shown in Fig. 2(d), all six model–source interactions are negative in the first construction, with confidence intervals entirely below zero. An independent construction reproduces five of the six effects, with only the NVILA object interval crossing zero. Further matched-interaction analyses and competence checks appear in **Appendix** B.5. These results establish non-additive source evidence under concurrent motion without isolating a failure to compose otherwise mastered single-source judgments.

Implications. The three findings expose a failure in motion-source reasoning. Motion evidence is not always bound to the queried entity, recoverable source information does not reliably influence decisions, and source-specific evidence becomes non-additive under concurrent motion. Improving motion-source reasoning therefore requires more than stronger temporal features: motion evidence must be bound to the query, used selectively by source, and remain composable under concurrent motion. These insights motivate the source-aware design introduced next.

4 MoSDeR: Motion-Source Decomposition and Reasoning

4.1 Framework Overview

As illustrated in Fig. 3, MoSDeR augments a pretrained VLM with a structured pipeline for source binding, selective routing, and joint composition. Given a video V and a language query q, the visual encoder produces tokens $X = E_{\text{vis}}(V)$, which the language decoder consumes to generate an answer $\hat{a}$. MoSDeR preserves this native multimodal interface while introducing two query-conditioned factors that organize temporal evidence according to its physical source.

The framework contains three coupled components. First, *query-conditioned source binding* associates visual evidence with the queried entity and derives object and camera factors S_o and S_c from target-related and contextual dynamics. Second, *source-selective adaptation* connects each factor to both visual tokens and a dedicated decoder branch, allowing structured source evidence to influence language decisions directly. Third, *non-exclusive source composition* retains both pathways as distinct but jointly available inputs to the decoder when camera and object motion coexist. The following subsections describe each component together with the supervision used to establish its intended behavior during training and inference in a unified model.

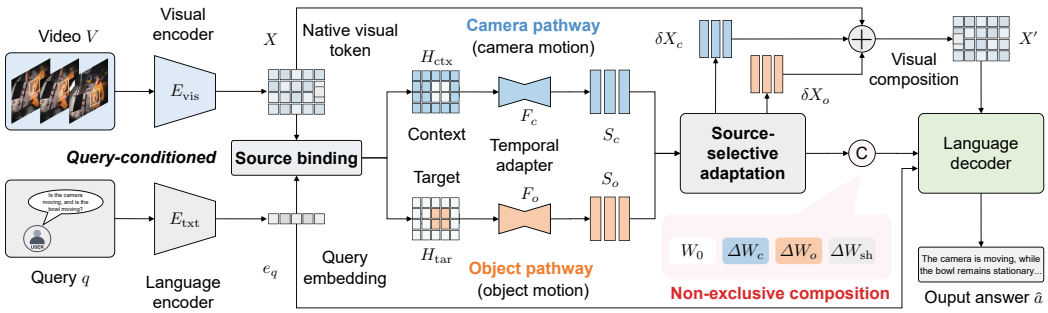


Figure 3: **Overview of MoSDeR.** Given a video and a query, MoSDeR first performs query-conditioned source binding to partition temporal evidence into target-related and context-related representations, from which object- and camera-motion factors are learned. Each factor is selectively routed to visual residual updates and decoder weight adaptation. Finally, both source-specific updates are non-exclusively composed within the native VLM decoder for joint reasoning.

4.2 Query-Conditioned Source Binding

Design Rationale. Finding 1 shows that motion recognition alone does not ensure that the resulting judgment follows the entity specified by the query. A scene-level temporal representation is therefore insufficient when the same video can support different object-motion judgments under different target queries. MoSDeR first binds visual evidence to the queried entity and then factorizes target-related and contextual dynamics into object- and camera-oriented temporal representations.

Query-Conditioned Evidence Allocation. Let $X \in \mathbb{R}^{T\times N\times D}$ denote the native visual tokens and $M_q \in [0,1]^{T\times N}$ their soft assignment to the query target. The resulting query-conditioned pooling operation produces temporally indexed target and context features $H_{\text{tar}} = \text{Pool}(X, M_q)$ and $H_{\text{ctx}} = \text{Pool}(X, 1 - M_q)$. This partition allocates evidence without assigning its source: target regions may contain viewpoint changes, while context need not remain static.

During source-supervised learning, available target boxes instantiate M_q; for general video questions, we predict it from the visual tokens and question stem. From the frozen question embedding e_q, we form $k_q = W_q\,\text{LN}(e_q)$ and $k_{tij} = W_v\,\text{LN}(X_{tij})$. Their scaled dot product and a tokenwise bias determine each query-conditioned target-membership score:

$$M_{q,tij} = \sigma\left(\frac{k_{tij}^{\mathsf{T}} k_q}{\sqrt{d_k}} + w_b^{\mathsf{T}} k_{tij} + b\right). \tag{1}$$

The allocator is trained with balanced binary cross-entropy on sparse target-box occupancy annotations and is then frozen for downstream adaptation in all later training stages.

Source-Specific Temporal Factorization. The allocated streams serve as inputs to object-oriented and camera-oriented temporal adapters rather than constituting source assignments by themselves. Given $H = \{H_t\}_{t=1}^{T}$, adapter $s \in \{o, c\}$ first computes $U_{s,t} = \text{LN}_{s,\text{in}}(H_t)$ and the temporal difference $d_{s,t} = U_{s,t} - U_{s,t-1}$, with $d_{s,1} = 0$. The implemented two-layer GELU bottleneck defines the source-specific projection as

$$P_s(x) = W_s^{\text{up}}\,\text{GELU}\big(W_s^{\text{down}} x + b_s^{\text{down}}\big) + b_s^{\text{up}}. \tag{2}$$

The adapter then mixes the projected features across the temporal dimension:

$$Z_s = P_s([U_s; d_s]) + E_s, \qquad F_s(H) = \text{LN}_{s,\text{out}}(Z_s + \text{GELU}(T_s(Z_s)))\,, \tag{3}$$

where E_s is a temporal embedding and T_s mixes information across time. Factors $S_o = F_o(H_{\text{tar}})$ and $S_c = F_c(H_{\text{ctx}})$ share temporal indexing but use distinct transformations.

4.3 Source-Selective Adaptation

Design Rationale. Finding 2 reveals that source information can be recoverable from intermediate representations without reliably influencing the corresponding answer. Explicit factors S_o and S_c are therefore useful only if they are connected to the native decision pathway. MoSDeR routes each factor into both the visual representation and source-specific decoder adaptation.

Source-Conditioned Visual and Decoder Updates. Let $M_o = M_q$, $M_c = 1 - M_q$, and let $\mathcal{E}(\cdot)$ align a source sequence with the native visual time axis and broadcast it spatially. Before residual

refinement, source s produces the spatially allocated residual update $\delta X_s^{(0)} = M_s \odot \mathcal{E}(\eta S_s)$, where η injects source-structured temporal evidence without composing the two routes.

Inside selected decoder projections, each source factor modulates a dedicated low-rank branch (Hu et al., 2022). The corresponding gate and source-conditioned low-rank update are defined as

$$g_s = \tanh(G_s \operatorname{mean}_t S_s), \qquad \delta_s(h; S_s) = \alpha B_s[(A_s h) \odot g_s]. \tag{4}$$

Here, A_s and B_s are source-specific low-rank factors and α is their scaling coefficient; together, the visual and decoder interfaces let each factor affect generation through complementary routes.

Source-Factor Grounding. We ground the two pathways using trajectory-derived binary source labels y_c and y_o, but apply supervision through the native decoder rather than an auxiliary classifier. Let n, c, o, and b denote the four canonical answers, and let $\ell_s(k)$ be the mean answer-token log probability of candidate k under source route s. The grounding margins are

$$m_c = \ell_c(c) - \ell_c(n), \qquad m_o = \frac{1}{2}\left[\ell_o(o) - \ell_o(n) + \ell_o(b) - \ell_o(c)\right]. \tag{5}$$

This asymmetry reflects the implemented supervision sets: camera grounding uses direct camera-only versus neither pairs, whereas object grounding preserves object contrast across both camera states by averaging the two available toggles. We retain these implemented contrasts rather than introducing an untested symmetric objective and optimize the factorization loss

$$\mathcal{L}_F = \sum_{s \in \{c,o\}} \operatorname{softplus}(-(2y_s - 1)(m_s + b_s)), \tag{6}$$

where b_s is a learned intercept. Only the supervised source pathway is active at this stage, grounding its temporal adapter, decoder branch, and visual injection.

4.4 Non-Exclusive Source Composition

Design Rationale. Finding 3 shows that source-specific evidence becomes less effective when camera and object motion co-occur. This motivates preserving the identity of both sources while allowing their contributions to coexist during generation. MoSDeR retains camera and object motion as two non-exclusive factors rather than using a single categorical source label.

Joint Source Composition. To preserve the native visual stream without suppressing either route, we combine the two source-conditioned visual updates as $X^{(0)} = X + \delta X_o^{(0)} + \delta X_c^{(0)}$. At each selected decoder projection, the shared and source-specific routes compose as

$$h' = W_0 h + \delta_o(h; S_o) + \delta_c(h; S_c) + \delta_{\mathrm{sh}}(h), \tag{7}$$

where W_0 is frozen and δ_{sh} is a shared language adaptation. Because the camera and object routes are conditioned independently, activating one source does not suppress the other. Their joint binary state yields four motion configurations: neither source moves, only the camera moves, only the object moves, or both sources move. This avoids a separate four-way head while keeping both source routes active within a single decoder pass for concurrent-motion reasoning.

Language Alignment and Refinement. After source-factor grounding, we freeze the source components and train the shared language branch to produce complete answers. With token sets I_C, I_O, I_S, and I_D for the four output fields, we optimize the following field-balanced generation objective:

$$\mathcal{L}_G = \frac{1}{4} \sum_{k \in \{C,O,S,D\}} \mathrm{NLL}_{I_k}, \tag{8}$$

The field-balanced loss prevents token-count dominance, and we refine the visual route as

$$\Delta_s = \beta\, \sigma\big(\gamma_s(\bar{S}_s)\big) \odot \tanh\big(R_s(\bar{S}_s)\big), \qquad \bar{S}_s = \mathrm{LN}(\mathrm{stopgrad}(S_s)). \tag{9}$$

The refined source updates enter the final visual route through the following additive composition:

$$\delta X_s = M_s \odot \mathcal{E}(\eta S_s + \Delta_s), \qquad X' = X + \delta X_o + \delta X_c. \tag{10}$$

At this refinement stage, all preceding parameters remain fixed, and only the residual functions γ_s and R_s are optimized. Zero initialization of the residual output preserves the preceding solution at the start of refinement, while β bounds each correction throughout optimization.

4.5 Training and Inference

Downstream Adaptation. During downstream adaptation, the learned query-conditioned allocator replaces externally supplied target boxes. The training objective jointly optimizes general video-language QA and source-supervised rehearsal in a unified loss:

$$\mathcal{L}_{\mathrm{QA}} = \mathbb{E}_{(V,q,a)\sim\mathcal{D}_{\mathrm{QA}}}\ell(a \mid V, q) + \mathbb{E}_{(V,B,q_{\mathrm{src}},a)\sim\mathcal{D}_{\mathrm{src}}}\ell(a \mid V, B, q_{\mathrm{src}}), \tag{11}$$

where ℓ is the mean answer-token negative log-likelihood. Under this objective, the temporal source modules, source-conditioned decoder factors and gates, shared language branch, and explicit residual modules are jointly updated. The base VLM, query-conditioned allocator, and factor intercepts remain frozen. The rehearsal term preserves the learned camera and object source structure while the model adapts to broader video-language tasks under the unified training objective.

Inference. For general video-language reasoning, MoSDeR receives only the video V and language query q. The query-conditioned allocator produces M_q, the temporal pathways derive S_o and S_c, and their source-conditioned contributions are composed directly by the native decoder. No trajectories, source labels, or intermediate motion states are required. For controlled MoSA evaluation, target boxes instantiate M_q to separate attribution from localization; they identify the queried target without supplying trajectories, motion-source labels, or answers to the model.

5 Experiments

5.1 Experimental Setting

Datasets. To evaluate motion-source attribution, we curate **MoSA** (see **Appendix** Fig. A1), a target-level benchmark spanning egocentric, allocentric, indoor-interaction, and autonomous-driving settings. The details are provided in **Appendix** A. We also evaluate on OmniVCHall (Xing et al., 2026), an independent video-hallucination benchmark with 9,027 VQA samples from 823 videos. It covers eight hallucination types, including object, action, relation, temporal, and camera-related reasoning, and evaluates isolated and compositional hallucinations. We use it to assess whether MoSDeR generalizes beyond motion-source attribution to broader video-language reasoning.

Metrics. For MoSA, we report accuracy for four motion states (Neither, Camera, Object, and Both), followed by source-dataset, macro-averaged, and pooled results. For OmniVCHall (Xing et al., 2026), we follow the official protocol and report Isolated, Compositional, and Overall accuracy.

Implementation Details. Fine-tuning uses backbone-specific LoRA (Hu et al., 2022), whereas MoSDeR retains native visual tokenization and decoding. Each source pathway uses a two-layer GELU bottleneck, 10-to-10 temporal mixer, low-rank decoder branch, and zero-initialized bounded residual. Following prior work (Yang et al., 2026; Cho et al., 2026), all backbones sample 20 frames at native resolutions; we use bottleneck width 32, decoder rank 8, effective batch size 8, and 625 updates per stage. AdamW learning rates are 10^{-4}/2×10^{-4}/10^{-2} for temporal adapters/decoder branches/decision biases during source factorization, 10^{-4} for shared adaptation, and 5×10^{-4} for residual refinement. After occupancy supervision, we freeze the allocator; refinement updates only γ_s and R_s. Downstream adaptation jointly updates the temporal, decoder, shared-language, and residual modules with source-supervised rehearsal. For fair comparison, trained methods use the same raw backbone and ordered MoSA split. Complete details appear in **Appendix** C.

5.2 Results and Analysis

Motion-Source Attribution on MoSA. MoSA tests whether VLMs attribute motion to its physical source. On the test set in Tab. 1, MoSDeR improves Overall over budget-matched fine-tuning by 11.08%, 11.75%, and 16.25% on NVILA, Qwen3, and Molmo2, respectively. Gains are largest on Both (48.00%, 26.00%, and 36.67%), supporting simultaneous preservation of camera and object evidence; Object also improves by 13.67–18.67%. Consistently, the fixed 400-case validation audit in Fig. 4 is more diagonal after MoSDeR and shows fewer cross-state confusions than raw and fine-tuned models, indicating improved source separation rather than a dominant-state bias. TACO saturates in 11 of 12 trained configurations, so we emphasize the three non-saturated sources: MoSDeR consistently outperforms budget-matched fine-tuning on ADT and AV2 across all three

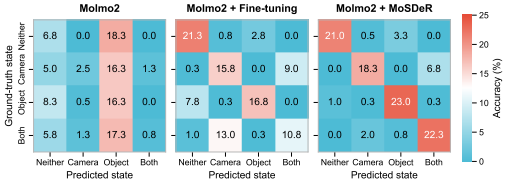


Figure 4: **Four-state confusion matrices.**

Table 1: **Motion-source attribution on MoSA.** Aggregate attribution report mixed-QA answer accuracy within the corresponding ground-truth motion-state groups. Avg. macro-averages source datasets, while Overall pools all mixed-QA samples. Best Overall results are bolded.

Model / Method	Aggregate attribution				Source dataset					Overall
	Neither	Camera	Object	Both	ADT	HOT3D	TACO	AV2	Avg.	
NVILA	1.67	32.67	63.67	85.33	30.63	46.97	66.09	53.54	49.31	45.83
NVILA + Fine-tuning	72.33	90.67	63.33	36.33	58.75	69.70	100.00	59.58	72.01	65.67
NVILA + 4DP-QA (Cho et al., 2026)	64.33	78.67	70.33	79.33	72.50	74.24	80.46	71.04	74.56	73.17
NVILA + 4D-RGPT (Yang et al., 2026)	71.67	89.00	63.67	37.00	59.17	69.70	100.00	58.33	71.80	65.33
NVILA + MoSDeR (Ours)	73.67	72.00	77.00	84.33	80.00	75.76	100.00	65.21	80.24	**76.75**
Qwen3	54.67	43.67	87.00	48.67	48.96	42.42	90.23	58.75	60.09	58.50
Qwen3 + Fine-tuning	85.00	73.67	75.33	58.33	71.88	77.27	100.00	63.96	78.28	73.08
Qwen3 + 4DP-QA (Cho et al., 2026)	95.00	68.33	93.33	70.33	87.29	69.70	100.00	71.25	82.06	81.75
Qwen3 + 4D-RGPT (Yang et al., 2026)	84.00	71.67	75.33	60.00	71.04	81.82	100.00	63.33	79.05	72.75
Qwen3 + MoSDeR (Ours)	81.67	84.00	89.33	84.33	91.46	75.76	100.00	73.96	85.29	**84.83**
Molmo2	58.67	37.67	62.00	30.00	50.83	39.39	62.64	38.75	47.91	47.08
Molmo2 + Fine-tuning	87.00	75.00	73.33	52.00	71.04	74.24	100.00	62.08	76.84	71.83
Molmo2 + 4DP-QA (Cho et al., 2026)	92.67	65.00	99.00	78.67	90.42	71.21	100.00	73.13	83.69	83.83
Molmo2 + 4D-RGPT (Yang et al., 2026)	86.00	76.00	73.33	50.67	71.88	72.73	100.00	60.62	76.31	71.50
Molmo2 + MoSDeR (Ours)	89.33	82.33	92.00	88.67	95.83	80.30	100.00	77.08	88.30	**88.08**

backbones. Within the AV2 subset, where all four states co-occur in one source, MoSDeR also attains the highest pooled and Both accuracy (see **Appendix** Tab. A11), indicating that the gains are not solely attributable to cross-dataset source identity.

Downstream Video-Language Reasoning. We evaluate on OmniVCHall (Xing et al., 2026) to assess whether MoSDeR transfers beyond MoSA without target boxes at inference. As shown in Tab. 2, MoSDeR achieves the best Overall accuracy for all three backbones, averaging 74.07%, outperforming raw VLMs, budget-matched fine-tuning, and TriCD by 13.58%, 7.47%, and 7.20%, respectively. Compared with fine-tuning, Overall accuracy improves by 10.04%, 0.11%, and 12.27% on NVILA, Qwen3, and Molmo2, while Compositional accuracy improves consistently by 0.72–12.37%. Isolated accuracy also improves on NVILA and Molmo2, with only a slight decrease on Qwen3 (-0.28%). These results demonstrate that MoSDeR generalizes to broader video-language reasoning. Representative corrected predictions are shown in **Appendix** Fig. A7.

Table 2: **Downstream evaluation on OmniVCHall.** YNQA and MCQA denote yes/no and multiple-choice question answering, respectively. Each Overall pools its corresponding columns.

Model / Method	Isolated			Compositional			Overall
	YNQA	MCQA	Overall	YNQA	MCQA	Overall	
NVILA	60.60	51.50	56.90	58.45	41.99	51.80	54.89
NVILA + Fine-tuning	58.37	61.26	59.54	68.72	67.48	68.22	62.97
NVILA + TriCD (Xing et al., 2026)	68.84	61.73	65.95	67.52	58.16	63.73	65.08
NVILA + MoSDeR (Ours)	**69.15**	**70.90**	**69.86**	**76.81**	**79.36**	**77.84**	**73.01**
Qwen3	73.89	60.28	68.36	77.54	61.57	71.08	69.43
Qwen3 + Fine-tuning	**79.43**	64.20	**73.24**	80.43	65.84	74.53	73.75
Qwen3 + TriCD (Xing et al., 2026)	72.63	63.34	68.85	**82.04**	**71.57**	**77.81**	72.39
Qwen3 + MoSDeR (Ours)	78.01	**65.59**	72.96	79.71	68.68	75.25	**73.86**
Molmo2	62.03	50.12	57.18	64.49	46.26	57.12	57.16
Molmo2 + Fine-tuning	69.30	52.42	62.44	72.22	51.96	64.03	63.07
Molmo2 + TriCD (Xing et al., 2026)	65.07	55.12	61.02	72.54	57.26	66.36	63.13
Molmo2 + MoSDeR (Ours)	**74.53**	**74.83**	**74.65**	**74.64**	**79.00**	**76.40**	**75.34**

Ablation Study. Tab. 3 evaluates simpler baselines and components on the MoSA validation set using Molmo2. MoSDeR reaches 89.17% Overall versus 75.08% for fine-tuning. Adding camera/object auxiliary supervision raises shared LoRA to 77.83%, but Both remains 56.33%, so supervision alone does not recover MoSDeR. Global binding lowers Overall to 66.50%, especially for Object and Both, underscoring query-target evidence binding. Shared factorization keeps the same M_q binding but uses a single shared factor, reaching 84.17%; the 5.00-point gap (9.33 on Both) reflects distinct source pathways rather than binding alone. Suppressing source-specific de-

Table 3: **Ablation on MoSA validation set.**

Method	Neither	Camera	Object	Both	Overall
Fine-tuning	85.00	77.67	78.33	59.33	75.08
Source-supervised Fine-tuning	97.00	82.00	76.00	56.33	77.83
Global Binding	97.67	94.67	36.67	37.00	66.50
Shared Factorization	73.67	87.33	95.00	80.67	84.17
w/o Visual Injection	83.67	84.67	95.67	85.00	87.25
w/o Decoder Routing	75.33	25.33	57.33	13.00	42.75
Exclusive Composition	93.33	96.67	38.67	40.33	67.25
MoSDeR	87.00	83.67	96.00	90.00	**89.17**

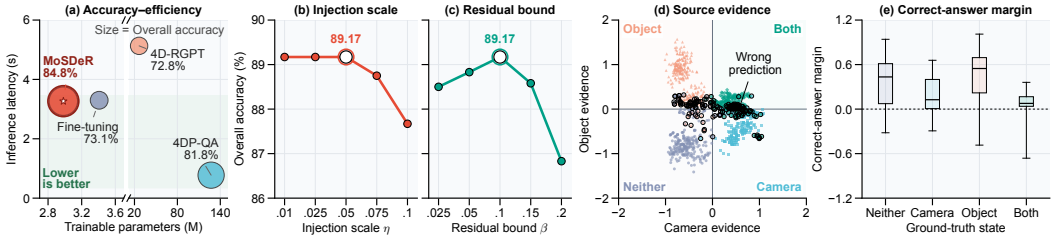
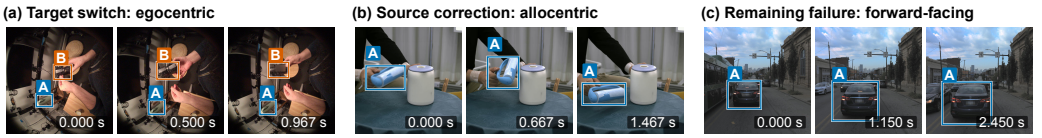


Figure 5: **Diagnostics.** (a) Accuracy–efficiency trade-off. (b)–(c) Molmo2 sensitivity to η and β. (d)–(e) Source-score contrasts and correct-answer margins on 1,200 state-balanced samples.



(a) A: Background item	
GT	Camera moving; object stationary
Raw	Camera stationary; object moving
Ours	Camera moving; object stationary

(a) B: Handheld item	
GT	Camera moving; object moving
Raw	Camera stationary; object moving
Ours	Camera moving; object moving

(b) A: Blue cylindrical tool	
GT	Camera stationary; object moving
Raw	Camera moving; object stationary
Ours	Camera stationary; object moving

(c) A: Black car ahead	
GT	Camera moving; object stationary
Raw	Camera stationary; object moving
Ours	Camera moving; object moving

Figure 6: **Qualitative motion-source attribution on MoSA.**

coder adapters reduces Overall and Both accuracy to 42.75% and 13.00%, whereas removing visual injection retains 87.25% Overall, indicating unequal contributions from the two pathways. Exclusive composition reduces Overall accuracy to 67.25%, with Object and Both falling to 38.67% and 40.33%, supporting the need to preserve camera and object evidence as non-exclusive factors. Overall, these results validate the complementary contributions of the components in MoSDeR.

Computational Efficiency. Fig. 5(a) illustrates the accuracy–efficiency trade-off on Qwen3, with full results for all backbones given in **Appendix** Tab. A12. MoSDeR attains the best Overall accuracy (84.8%) using only 2.982M trainable parameters and 3.267s inference latency, closely matching budget-matched fine-tuning (3.408M and 3.298s). Although 4DP-QA is faster, it requires 42.6× more trainable parameters; 4D-RGPT uses 8.9× more parameters and is also slower. These results show that MoSDeR offers the most favorable accuracy–efficiency trade-off on Qwen3.

Hyperparameter Sensitivity. Figs. 5(b) and 5(c) demonstrate broad robustness: Overall remains 89.17% for $\eta \in [0.01, 0.05]$ and within 0.67% of the selected result for $\beta \in [0.025, 0.15]$. This stable plateau confirms MoSDeR's robustness to moderate parameter variation without narrow tuning. White-centered markers denote the selected settings, with complete sweeps in **Appendix** Tab. A13.

Attribution Analysis. Together, Figs. 5(d), 5(e) and 6 connect score separation to observable source attribution. In Fig. 5(d), the four states occupy their expected camera–object regions, whereas errors cluster near zero-evidence boundaries, localizing failures to weakly separated source contrasts. In Fig. 5(e), 256 of 300 Both samples have positive correct-answer margins; camera-only is the strongest incorrect alternative for 233 of them, identifying concurrent motion as the closest remaining ambiguity. Fig. 6 grounds these patterns: target switching preserves camera attribution while changing object attribution, source routing corrects an object-motion error under a fixed camera, and non-exclusive composition recovers the missing camera cue when both sources move. The resulting chain runs from query-conditioned binding through source separation to composed answers; the broader count-based audit across all four source states appears in **Appendix** Tab. A14.

6 Conclusion and Discussion

We identified **motion-source misattribution** as a recurring failure mode in video-language models: temporal evidence may be present yet remain poorly bound to the queried target and physical source. **MoSDeR** addresses this gap through query-conditioned binding, source-specific temporal factorization, source-selective adaptation, and non-exclusive composition of camera and object evidence. Across three VLM families, it consistently improves motion-source attribution, especially under concurrent motion, while gains on OmniVCHall indicate broader benefits for video-language reasoning. Our current study focuses on rigid targets and controlled source attribution, leaving non-rigid motion and broader source configurations for future work. These results motivate explicitly organizing temporal evidence by its physical source for more reliable video-language reasoning.

Reproducibility statement

Dataset construction and splits are documented in **Appendix** A and **Appendix** Tab. A5, evaluation controls in **Appendix** Tab. A6, and model configurations in **Appendix** C.4 and **Appendix** Tab. A10. Source code is included in the supplementary material for examination and reproduction.

Ethics statement

MoSA uses released public datasets under their terms and involves no new human-subject collection. Resources are for research; deployment should consider privacy, surveillance, and domain bias.

AI Use Statement

AI-assisted tools were used only for improving writing clarity, grammar checking, and minor editing during manuscript preparation. The authors are solely responsible for the content, including the research methodology, experimental results, and conclusions.

References

Kyungho Bae, Jinhyung Kim, Sihaeng Lee, Soonyoung Lee, Gunhee Lee, and Jinwoo Choi. MASH-VLM: Mitigating action-scene hallucination in video-LLMs through disentangled spatial-temporal representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13744–13753, 2025.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Prithviraj Banerjee, Sindi Shkodrani, Pierre Moulon, Shreyas Hampali, Shangchen Han, Fan Zhang, Linguang Zhang, Jade Fountain, Edward Miller, Selen Basol, Richard Newcombe, Robert Wang, Jakob Julian Engel, and Tomáš Hodaň. HOT3D: Hand and object tracking in 3D from egocentric multi-view videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7061–7071, 2025.

Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is Space-Time attention all you need for video understanding? In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 813–824, 2021.

Seokju Cho, Abhishek Badki, Hang Su, Jindong Jiang, Ziyao Zeng, Seungryong Kim, Sifei Liu, and Orazio Gallo. 4DP-QA: Scalable QA for 4D perception in vision language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23869–23879, 2026.

Christopher Clark, Jieyu Zhang, Zixian Ma, Jae Sung Park, Rohun Tripathi, Sangho Lee, Mohammadreza Salehi, Jason Ren, Chris Dongjoo Kim, Yinuo Yang, Vincent Shao, Yue Yang, Weikai Huang, Ziqi Gao, Taira Anderson, Jianrui Zhang, Jitesh Jain, George Stoica, Ali Farhadi, and Ranjay Krishna. Molmo2: Open weights and data for vision-language models with video understanding and grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28652–28668, 2026.

Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. Amnesic probing: Behavioral explanation with amnesic counterfactuals. *Transactions of the Association for Computational Linguistics*, 9:160–175, 2021.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, Peixian Chen, Yanwei Li, Shaohui Lin, Sirui Zhao, Ke Li,

Tong Xu, Xiawu Zheng, Enhong Chen, Caifeng Shan, Ran He, and Xing Sun. 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*, pp. 24108–24118, 2025.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2733–2743, 2019.

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 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8450–8460, 2025.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Chaoyu Li, Eun Woo Im, and Pooyan Fazli. VidHalluc: Evaluating temporal hallucinations in multimodal large language models for video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13723–13733, 2025.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, Limin Wang, and Yu Qiao. MVBench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22195–22206, 2024.

Xiang Li, Heqian Qiu, Lanxiao Wang, Hanwen Zhang, Chenghao Qi, Linfeng Han, Huiyu Xiong, and Hongliang Li. Challenges and trends in egocentric vision: A survey. *Machine Intelligence Research*, 23(1):1–33, 2026.

Kevin Qinghong Lin, Jinpeng Wang, Mattia Soldan, Michael Wray, Rui Yan, Eric Z. Xu, Difei Gao, Rong-Cheng Tu, Wenzhe Zhao, Weijie Kong, Chengfei Cai, Hongfa Wang, Dima Damen, Bernard Ghanem, Wei Liu, and Mike Zheng Shou. Egocentric video-language pretraining. In *Advances in Neural Information Processing Systems*, volume 35, pp. 7575–7586, 2022.

Zhiqiu Lin, Siyuan Cen, Daniel Jiang, Jay Karhade, Hewei Wang, Chancharik Mitra, Yu Tong Tiffany Ling, Yuhan Huang, Sifan Liu, Mingyu Chen, Rushikesh Zawar, Xue Bai, Yilun Du, Chuang Gan, and Deva Ramanan. Towards understanding camera motions in any video. In *Advances in Neural Information Processing Systems*, volume 38, pp. 148814–148857, 2025.

Yun Liu, Haolin Yang, Xu Si, Ling Liu, Zipeng Li, Yuxiang Zhang, Yebin Liu, and Li Yi. TACO: Benchmarking generalizable bimanual Tool-ACtion-Object understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21740–21751, 2024.

Zhijian Liu, Ligeng Zhu, Baifeng Shi, Zhuoyang Zhang, Yuming Lou, Shang Yang, Haocheng Xi, Shiyi Cao, Yuxian Gu, Dacheng Li, Xiuyu Li, Haotian Tang, Yunhao Fang, Yukang Chen, Cheng-Yu Hsieh, De-An Huang, An-Chieh Cheng, Jinyi Hu, Sifei Liu, Ranjay Krishna, Pavlo Molchanov, Jan Kautz, Hongxu Yin, Song Han, and Yao Lu. NVILA: Efficient frontier visual language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4122–4134, 2025.

Zijie Meng. OrthoMotion: Disentangling camera and subject motion via geometry–semantics orthogonal attention. *arXiv preprint arXiv:2606.22835*, 2026.

Xiaqing Pan, Nicholas Charron, Yongqian Yang, Scott Peters, Thomas Whelan, Chen Kong, Omkar Parkhi, Richard Newcombe, and Carl Yuheng Ren. Aria digital twin: A new benchmark dataset for egocentric 3D machine perception. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 20133–20143, 2023.

Viorica Pătrăucean, Lucas Smaira, Ankush Gupta, Adria Recasens, Larisa Markeeva, Dylan Banarse, Skanda Koppula, Joseph Heyward, Mateusz Malinowski, Yi Yang, Carl Doersch, Tatiana Matejovicova, Yury Sulsky, Antoine Miech, Alexandre Fréchette, Hanna Klimczak, Raphael Koster, Junlin Zhang, Stephanie Winkler, Yusuf Aytar, Simon Osindero, Dima Damen, Andrew Zisserman, and João Carreira. Perception test: A diagnostic benchmark for multimodal video models. In *Advances in Neural Information Processing Systems*, volume 36, pp. 42748–42761, 2023.

Shraman Pramanick, Yale Song, Sayan Nag, Kevin Qinghong Lin, Hardik Shah, Mike Zheng Shou, Rama Chellappa, and Pengchuan Zhang. EgoVLPv2: Egocentric video-language pre-training with fusion in the backbone. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5285–5297, 2023.

Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *Advances in Neural Information Processing Systems*, volume 27, pp. 568–576, 2014.

Chinthani Sugandhika, Chen Li, Deepu Rajan, and Basura Fernando. Know-Show: Benchmarking video-language models on spatio-temporal grounded reasoning. *arXiv preprint arXiv:2512.05513*, 2025.

Yunlong Tang, Jing Bi, Siting Xu, Luchuan Song, Susan Liang, Teng Wang, Daoan Zhang, Jie An, Jingyang Lin, Rongyi Zhu, Ali Vosoughi, Chao Huang, Zeliang Zhang, Pinxin Liu, Mingqian Feng, Feng Zheng, Jean-Luc Gaudiot, Ping Luo, Jiebo Luo, and Chenliang Xu. Video understanding with large language models: A survey. *IEEE Transactions on Circuits and Systems for Video Technology*, 36(2):1355–1376, 2026.

Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In *Advances in Neural Information Processing Systems*, volume 35, pp. 10078–10093, 2022.

Heng Wang and Cordelia Schmid. Action recognition with improved trajectories. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3551–3558, 2013.

Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Zun Wang, Yansong Shi, Tianxiang Jiang, Songze Li, Jilan Xu, Hongjie Zhang, Yifei Huang, Yu Qiao, Yali Wang, and Limin Wang. InternVideo2: Scaling foundation models for multimodal video understanding. In *European Conference on Computer Vision*, pp. 396–416, 2024a.

Yuxuan Wang, Yueqian Wang, Dongyan Zhao, Cihang Xie, and Zilong Zheng. VideoHallucer: Evaluating intrinsic and extrinsic hallucinations in large video-language models. *arXiv preprint arXiv:2406.16338*, 2024b.

Benjamin Wilson, William Qi, Tanmay Agarwal, John Lambert, Jagjeet Singh, Siddhesh Khandelwal, Bowen Pan, Ratnesh Kumar, Andrew Hartnett, Jhony Kaesemodel Pontes, Deva Ramanan, Peter Carr, and James Hays. Argoverse 2: Next generation datasets for self-driving perception and forecasting. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks (NeurIPS Datasets and Benchmarks 2021)*, 2021.

Wenbin Xing, Quanxing Zha, Lizheng Zu, Mengran Li, Ming Li, and Junchi Yan. Learning to decode against compositional hallucination in video multimodal large language models. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026.

Zihui Xue, Kristen Grauman, Dima Damen, Andrew Zisserman, and Tengda Han. Seeing without pixels: Perception from camera trajectories. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 38836–38847, 2026.

Chiao-An Yang, Ryo Hachiuma, Sifei Liu, Subhashree Radhakrishnan, Raymond A. Yeh, Yu-Chiang Frank Wang, and Min-Hung Chen. 4D-RGPT: Toward region-level 4D understanding via perceptual distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 31042–31053, 2026.

Xiangyu Zeng, Kunchang Li, Chenting Wang, Xinhao Li, Tianxiang Jiang, Ziang Yan, Songze Li, Yansong Shi, Zhengrong Yue, Yi Wang, Yali Wang, Yu Qiao, and Limin Wang. TimeSuite: Improving MLLMs for long video understanding via grounded tuning. In *International Conference on Learning Representations*, 2025.

Jiacheng Zhang, Yang Jiao, Shaoxiang Chen, Na Zhao, Zhiyu Tan, Hao Li, Xingjun Ma, and Jingjing Chen. EventHallusion: Diagnosing event hallucinations in video LLMs. *arXiv preprint arXiv:2409.16597*, 2024.

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. LLaVA-Video: Video instruction tuning with synthetic data. *Transactions on Machine Learning Research*, 2025a.

Ziang Zhang, Zehan Wang, Guanghao Zhang, Weilong Dai, Yan Xia, Ziang Yan, Minjie Hong, and Zhou Zhao. DSI-Bench: A benchmark for dynamic spatial intelligence. *arXiv preprint arXiv:2510.18873*, 2025b.

Yue Zhao, Ishan Misra, Philipp Krähenbühl, and Rohit Girdhar. Learning video representations from large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6586–6597, 2023.

Chenhao Zheng, Jieyu Zhang, Mohammadreza Salehi, Ziqi Gao, Vishnu Iyengar, Norimasa Kobori, Quan Kong, and Ranjay Krishna. One trajectory, one token: Grounded video tokenization via panoptic sub-object trajectory. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 23156–23166, 2025.

Chenhao Zheng, Jieyu Zhang, Jianing Zhang, Weikai Huang, Ashutosh Kumar, Quan Kong, Oncel Tuzel, Chun-Liang Li, and Ranjay Krishna. TrajTok: Learning trajectory tokens enhances video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 31207–31214, 2026.

Shengchao Zhou, Yuxin Chen, Yuying Ge, Wei Huang, Jiehong Lin, Ying Shan, and Xiaojuan Qi. Learning to reason in 4D: Dynamic spatial understanding for vision language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9637–9646, 2026.

Orr Zohar, Xiaohan Wang, Yonatan Bitton, Idan Szpektor, and Serena Yeung-Levy. Video-STaR: Self-training enables video instruction tuning with any supervision. In *International Conference on Learning Representations*, 2025.

Appendix Contents

A Benchmark Construction **14**
A.1 Source Datasets . . . 14
A.2 Sample and Target Selection . . . 15
A.3 Motion-Source Annotation . . . 15
A.4 Query and Answer Construction . . . 16
A.5 Quality Control and Filtering . . . 16
A.6 Dataset Splits and Statistics . . . 17
A.7 Evaluation Protocol . . . 17

B Additional Empirical Analysis **18**
B.1 Diagnostic Data and Evaluation Protocol . . . 18
B.2 Qualitative AV2 Target-Switch Result . . . 18
B.3 Robustness of Partial Source-Information Recoverability . . . 18
B.4 Selective Effects of Internal Source Interventions . . . 19
B.5 Matched Motion Interactions and Interpretation Boundaries . . . 20

C Additional Implementation Details **22**
C.1 Backbone Instantiations . . . 22
C.2 Module Instantiation . . . 22
C.3 Stagewise Parameter Updates . . . 23
C.4 Reproducibility and Comparison Protocol . . . 23
C.5 Inference Interface . . . 23

D Additional Experimental Results **23**
D.1 Within-Source Motion-State Analysis . . . 23
D.2 Downstream Qualitative Analysis . . . 24
D.3 Computational Efficiency . . . 24
D.4 Hyperparameter Sensitivity . . . 25
D.5 Qualitative Case Audit on MoSA . . . 25

A Benchmark Construction

A.1 Source Datasets

We construct MoSA from selected recordings in Aria Digital Twin (ADT) (Pan et al., 2023), HOT3D (Banerjee et al., 2025), TACO V1 (Liu et al., 2024), and Argoverse 2 (AV2) (Wilson et al., 2021). As summarized in **Appendix** Fig. A1 and **Appendix** Tab. A1, these sources span complementary egocentric and allocentric viewpoints, indoor interactions, and vehicle-mounted outdoor scenes, with physical metadata distinguishing camera from target motion. Inclusion requires timestamp-compatible camera calibration or poses, rigid-target trajectories or poses, and an auditable target track throughout the sampled window under a consistent, source-independent eligibility procedure.

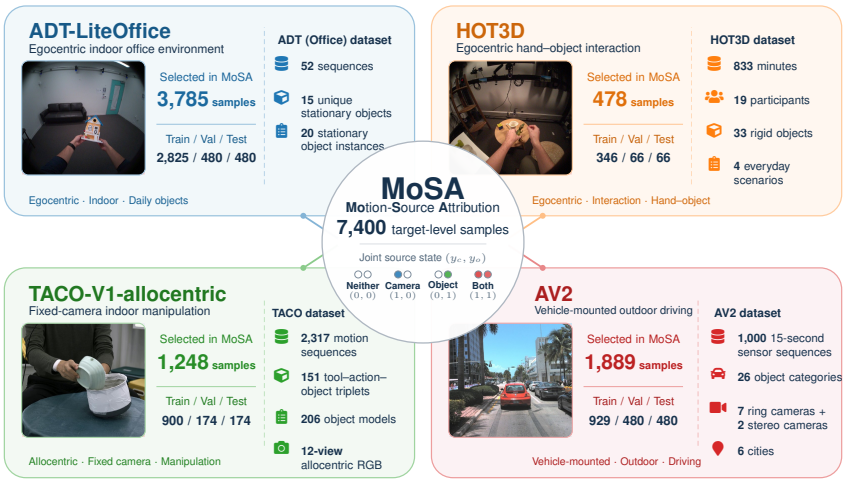


Figure A1: **Overview of the MoSA benchmark.** MoSA combines four complementary sources covering egocentric and allocentric views, indoor interaction and outdoor driving, and diverse camera and object motion configurations under a shared target-conditioned attribution protocol.

Table A1: **Source datasets and selected recording subsets.** The table identifies the viewpoints and traceable physical annotations used to construct the MoSA benchmark.

Dataset	Selected subset	Viewpoint	Physical annotations
ADT	Locally designated LiteOffice recognition sequences	Egocentric, indoor	Device trajectories, scene objects, instances, and boxes
HOT3D	Selected clips from RGB stream 214-1	Egocentric interaction, indoor	Camera and rigid-object world poses, boxes, and visibility
TACO V1	Valid manipulation windows from allocentric recordings	Fixed camera, indoor	Calibrated camera extrinsics, object poses, and target localization
AV2	Vehicle-mounted clips from `ring_front_center`	Vehicle-mounted, outdoor	Ego poses, sensor calibration, and world-referenced vehicle trajectories

A.2 Sample and Target Selection

Each sample corresponds to a single queried target and contains 20 temporally ordered RGB frames, timestamps, and a bounding-box track. Multiple targets within the same video window can yield separate samples. The sampled frames span approximately 1.5 s for ADT and TACO, 1.0 or 2.0 s for HOT3D, and 2.5 s for AV2. We include only targets with rigid-body annotations, excluding non-rigid people and animals as well as locally articulated cases. At inference, the supplied boxes specify the queried target; no ground-truth answers or motion-state labels are provided, and no visual markers are added to the frames.

A.3 Motion-Source Annotation

Camera and target-object motion are labeled as two non-exclusive binary factors relative to the fixed scene. Their combinations define four motion states: `neither`, `camera_only`, `object_only`, and `both`. Labels are derived and cached during preprocessing using source-specific native-window statistics. **Appendix** Tab. A2 summarizes the annotation criteria.

ADT, HOT3D, and AV2 use separate stationary and moving thresholds; intermediate cases are excluded to reduce ambiguity. ADT measures motion extent relative to the first valid pose. HOT3D uses accumulated path length divided by duration. TACO retains inherited `object_only` labels from valid manipulation windows under a fixed-camera constraint.

Exported positions use meters, and orientations use xyzw unit quaternions. ADT RGB-camera poses combine device trajectories with camera calibration, while motion labels retain the source world-device reference point. HOT3D uses native world poses. TACO camera poses are obtained by inverting the supplied world-to-camera extrinsics. AV2 combines ego poses with sensor calibration, using linear interpolation for position and spherical linear interpolation for orientation to align physical annotations with RGB timestamps at every selected video frame.

The exported 20-point pose sequences support temporal alignment and auditing, while MoSDeR uses the resulting states and language supervision throughout source-factor training.

Table A2: **Source-specific motion-label criteria.** Stationary and moving conditions are evaluated separately; intermediate cases are excluded. HOT3D and AV2 target labels use translation criteria. TACO retains inherited labels under fixed-camera valid-manipulation constraints.

Dataset	Factor	Stationary condition	Moving condition
ADT	Camera	Translation extent ≤ 0.02 m **and** rotation extent $\leq 2^\circ$	Translation extent ≥ 0.05 m **or** rotation extent $\geq 5^\circ$
	Object	Translation extent ≤ 0.01 m **and** rotation extent $\leq 1^\circ$	Translation extent ≥ 0.05 m **or** rotation extent $\geq 5^\circ$
HOT3D	Camera	Translation path/duration ≤ 0.005 m/s **and** rotation path/duration $\leq 0.5^\circ$/s	Translation path/duration ≥ 0.01 m/s **or** rotation path/duration $\geq 1^\circ$/s
	Object	World-translation path/duration ≤ 0.0025 m/s	World-translation path/duration ≥ 0.005 m/s
TACO V1	Camera	Fixed allocentric-camera constraint	Not included in the selected subset
	Object	Not included in the selected subset	Inherited object-only labels from valid manipulation windows
AV2	Camera	Net displacement ≤ 0.15 m **and** path length ≤ 0.25 m **and** rotation $\leq 2^\circ$	Net displacement ≥ 0.5 m **or** path length ≥ 0.75 m **or** rotation $\geq 5^\circ$
	Object	Net displacement ≤ 0.25 m **and** path length ≤ 0.4 m	Net displacement ≥ 1 m **or** path length ≥ 1.2 m

A.4 Query and Answer Construction

A fixed source-learning instruction requests Camera, Object, State, and Description: the first two encode binary factors, State their combination, and Description canonical supervision.

Each target-level sample receives one camera, object, or state-selection question. Camera/object questions use yes/no answers, while state questions use four alternatives; wording, option order, and answers derive deterministically from physical labels. Each validation/test split contains 400 questions per family and 100 per state–family combination by construction.

Fixed source-learning instruction.

```
Analyze the motion source in this 20-frame video for
the queried target specified by the target-box track.
Distinguish camera motion from target-object motion.
Respond using exactly four lines: Camera, Object, State,
Description.
```

Table A3: **Canonical source-state supervision.** Descriptions are fixed English supervision targets. Predictions are generated by the model rather than populated from ground-truth labels.

State	Camera	Object	Description
`neither`	`static`	`static`	the camera and queried target remain stationary.
`camera_only`	`moving`	`static`	the camera moves while the queried target remains stationary.
`object_only`	`static`	`moving`	the queried target moves while the camera remains stationary.
`both`	`moving`	`moving`	both the camera and queried target move.

A.5 Quality Control and Filtering

Quality control covers provenance, valid poses, temporal coverage, rigid targets, RGB decoding and order, box validity, visibility, and label ambiguity; source-specific criteria appear in **Appendix** Tab. A4. Functional tests verify data and evaluator consistency, including input invariance to ground-truth changes; audits combine metadata checks, sampled decoding, and targeted file inspection.

Table A4: **Source-specific quality-control criteria.** The scope column identifies the records to which each set of criteria applies during source-specific benchmark filtering and verification.

Scope	Verified criteria
Inherited ADT	At least 3 pose observations; maximum pose gap ≤ 0.15 s; box visibility ≥ 0.5; visible-frame fraction ≥ 0.7; box-area fraction ≥ 0.01; ambiguous rotation-only cases involving symmetric objects are excluded.
New ADT test	All 20 selected RGB frames require qualified target boxes; nearest box-to-image timestamp discrepancy ≤ 0.5 ms.
HOT3D	Predicted target visibility ≥ 0.7 in every original observation frame; mean amodal box-area fraction ≥ 0.003; in-image box fraction ≥ 0.95; consecutive source frames and target-pose coverage.

A.6 Dataset Splits and Statistics

MoSA contains 7,400 target-level samples divided into training, validation, and test partitions, with equal support for the four motion states within each split. **Appendix** Tab. A5 reports source counts, state balance, and the physical-window and acquisition-group totals. Multiple targets and overlapping windows are allowed within a split. Cross-split audits find no overlap in sample identities, physical windows, sequences, acquisition groups, ADT activity families, or exact frame content. Motion-state balance is enforced globally, with exactly 1,250/300/300 samples per state in the training/validation/test splits. Individual source datasets need not contain all four states because of their physical acquisition settings. Source identity is never provided as an explicit prompt field. Training uses iterative proportional fitting to target ADT/HOT3D/TACO/AV2 marginals of 35/15/15/35% and equal state marginals. Within each source–state group, physical windows receive equal total sampling mass; Tab. 1 reports source-wise results across all three backbones.

Table A5: **Dataset composition and sampling units.** Target-level sample counts are not counts of independent videos. Physical-window and acquisition-group totals are reported for each split.

Source or counting unit	Train	Validation	Test
ADT targets	2,825	480	480
HOT3D targets	346	66	66
TACO targets	900	174	174
AV2 targets	929	480	480
Total target-level samples	**5,000**	**1,200**	**1,200**
Samples per motion state	1,250	300	300
Distinct physical windows	3,577	943	647
Acquisition groups	96	40	34

A.7 Evaluation Protocol

Mixed-QA evaluation assigns one camera, object, or state-selection question to each target-level sample. All compared methods receive the same ordered 20 RGB frames, normalized target-box track, and assigned QA question with identical option order. Each method returns one yes/no answer or option letter, which is scored by the same answer normalization and correctness rule. **Appendix** Tab. A6 summarizes this matched evaluation protocol. Invalid outputs are counted as errors and reported separately. The release evaluator uses 1,000 acquisition-group bootstrap resamples stratified by source by default. Supplementary paired ablations use 2,000 source-weight-preserving group resamples with seed 20260917 for the reported uncertainty estimates.

Table A6: **Common mixed-QA evaluation configuration.** All methods receive identical external inputs and the same answer evaluator; methods differ only in how they process these inputs.

Component	Shared configuration for all compared methods
Video input	The same 20 ordered RGB frames with backbone-native image preprocessing.
Target specification	The same normalized xyxy box track identifies the queried target for every method.
Question input	The same assigned camera-motion, object-motion, or state-selection question, including option order.
Required answer	One yes/no answer or one option letter (A–D), according to the assigned question.
Scoring	The same answer normalization and exact correctness rule; invalid outputs count as errors.
Excluded information	Camera poses, object trajectories, source labels, and ground-truth motion states.

B Additional Empirical Analysis

B.1 Diagnostic Data and Evaluation Protocol

We distinguish camera motion relative to the scene from motion of the queried target relative to the scene; their binary combination defines neither, camera-only, object-only, and both. The natural-video diagnostics comprise 240 balanced queries from 127 AV2 sequences and 89 HOT3D queries from 32 physical windows, including 21 same-video target-switch pairs. AV2 supports state-balanced driving diagnostics, while HOT3D holds the video fixed across target changes to test camera-answer invariance and target-dependent judgments across both sources.

The two natural-data sources serve complementary roles: AV2 supplies the state-balanced driving diagnostic and the qualitative example in **Appendix** Fig. A2, whereas HOT3D supplies the 21 same-video target-switch pairs. Both also contribute to MoSA, aligning the motivating analysis with benchmark domains while retaining distinct diagnostic protocols.

Supplementary analyses separate probes of recoverability, native-margin interventions, and matched tests of cross-source interactions; four-state accuracy requires both source labels.

B.2 Qualitative AV2 Target-Switch Result

Appendix Fig. A2 shows the same AV2 video queried for two targets (Wilson et al., 2021). The visual evidence and camera state remain fixed while target identity and motion state change.

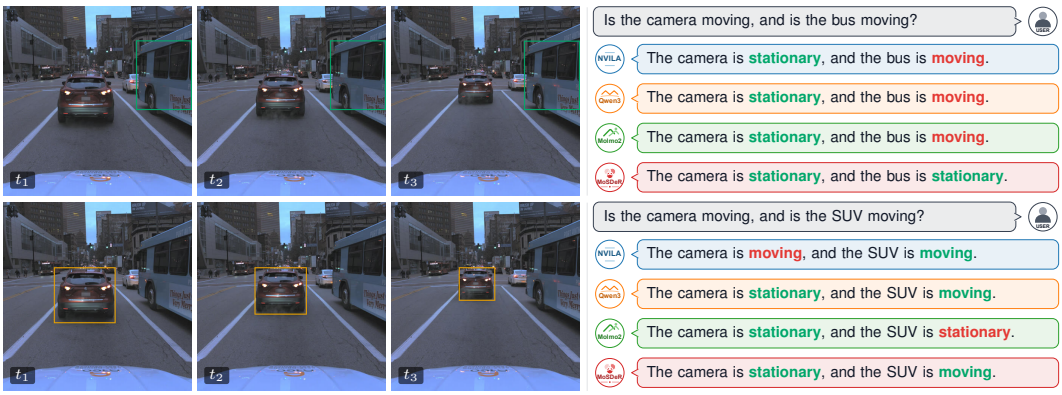


Figure A2: **Qualitative target-switch result on an AV2 video.** The two panels use the same AV2 frames but switch the queried target from the stationary bus (top) to the moving SUV (bottom); the camera remains stationary in both queries. Green text denotes a correct factor judgment and red text an incorrect judgment. In this example, every baseline makes a source-attribution error for at least one target, whereas MoSDeR correctly preserves the camera state and updates the target state after the query switch, matching both ground-truth source labels for both queries.

This pair exposes two query-binding failures. For the stationary bus, all three baselines preserve the camera judgment but misclassify target motion; with the moving SUV, Qwen3 recovers both factors, NVILA changes the camera judgment incorrectly, and Molmo2 fails to update the target judgment. MoSDeR instead preserves the camera state while changing the target state with the query.

B.3 Robustness of Partial Source-Information Recoverability

B.3.1 Holding Out Content Factors

To test whether recoverability depends on familiar content, we evaluate probes under separate hold-outs of target assets, camera-motion programs, target-motion programs, and backgrounds. Each condition withholds one family rather than all four simultaneously, and the probes predict camera and target motion rather than asset or background identity under every condition.

Across the three models and four holdouts, four-state accuracy ranges from 46.7% to 63.3%, above the 25% uniform reference (**Appendix** Fig. A3). Source information therefore remains partially recoverable when individual content-factor families are withheld from probe fitting.

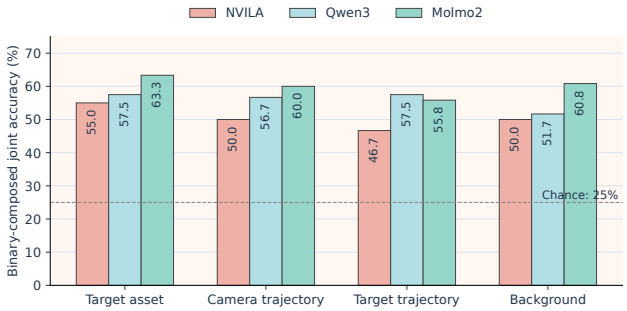


Figure A3: **Source-information recoverability under factor-family holdouts.** Each group withholds a target-asset, camera-trajectory, target-trajectory, or background family from probe fitting. Bars report four-state accuracy for probes on frozen model representations. Factors are held out separately. The dashed line marks the 25% uniform-guessing reference. Only point estimates are reported because uncertainty intervals were not measured for this diagnostic.

B.3.2 Simultaneous Content Isolation

A complementary strict-content diagnostic simultaneously separates scene assets, backgrounds, camera programs, object programs, and base scenes between training and held-out testing. Its 100 base units are split into 60 training, 20 validation, and 20 test units, with four motion conditions per unit. Thus, the test set contains 80 cases. Ridge regularization is selected using validation data from a fixed five-value grid. We additionally report the recorded reference interval from 100 label permutations rather than relying only on the nominal 25% chance level (**Appendix** Tab. A7).

Table A7: **Probe results under simultaneous content isolation.** Accuracy intervals resample base units, whereas the permutation interval summarises label-permuted controls rather than uncertainty in observed probe accuracy estimates for the evaluated models and held-out units.

Model	Four-state acc. (%) [95% CI]	Permutation mean (%)	Permutation reference interval (%)
NVILA	53.75 [48.75, 58.75]	24.36	[5.59, 42.50]
Qwen3	50.00 [50.00, 50.00]	28.39	[5.53, 55.06]
Molmo2	56.25 [51.25, 62.50]	25.84	[1.25, 47.56]

NVILA and Molmo2 achieve four-state accuracies above the upper endpoints of their recorded permutation reference intervals, whereas Qwen3's 50.0% score remains within its interval. Source information is therefore recoverable under strict content isolation for some models, although the evidence varies across families and supports partial rather than universal linear decodability.

B.4 Selective Effects of Internal Source Interventions

B.4.1 Intervention Construction and Measurement

We use matched counterfactual videos to construct additive source components at a specified language-model layer. Let h_{00}, h_{10}, h_{01}, and h_{11} denote hidden states for neither, camera-only, object-only, and both-motion conditions under the same question. The camera component is $h_{10} - h_{00}$, and the object component is $h_{01} - h_{00}$. To add object motion, we patch h_{10} with $h_{10} + (h_{01} - h_{00})$; to remove it, we patch h_{11} with $h_{11} - (h_{01} - h_{00})$. Camera interventions are defined analogously. These components come from matched hidden-state differences rather than fitted probe weights or parameters learned from diagnostic labels.

Interventions modify hidden activations during a forward pass. Model weights and the question are unchanged. Patching is restricted to the specified layer and token positions; subsequent computation proceeds normally. Because constructing a component uses matched counterfactual videos, this is a privileged diagnostic rather than a deployable video-only inference procedure.

We measure the first-answer-token margin, $\mathrm{logit}(\mathrm{Yes}) - \mathrm{logit}(\mathrm{No})$. For each relevant source, add and remove operations are signed according to the expected direction of change and averaged within a base unit. A random-direction control uses the same L_2 norm and patch scope. A second control measures the absolute margin change caused by interventions on the other source while holding the question fixed. The two contrasts are therefore the relevant signed effect minus the signed random-

control effect, and the relevant signed effect minus the absolute other-source effect. Selectivity requires the lower endpoints of both 95% bootstrap intervals to exceed zero. The diagnostic uses 30 base units; the intervals are computed from unit-level effects throughout the analysis.

Table A8: **Internal locations used for source interventions.** Each row pairs a relative language-model depth with the hidden-state token positions modified at that location.

Relative language-model depth	Hidden states modified
25.0%	All video-token positions
50.0%	All video-token positions
50.0%	The final input-token position
87.5%	The final input-token position

B.4.2 Layer Depth and Token Scope

Appendix Tab. A8 lists four internal locations. Relative depth indexes the language-model stack rather than video time. For a model with L layers, the zero-based layer is $\lceil \text{depth} \times L \rceil - 1$.

The final input token is the last position of the tokenized prompt before answer generation. These scopes identify internal activations; they do not mean replacing video frames or rewriting prompt text. The 50%-depth video-token condition is included in the full location scan rather than plotted again as a separate experiment. **Appendix** Fig. A4 reports the full scan.

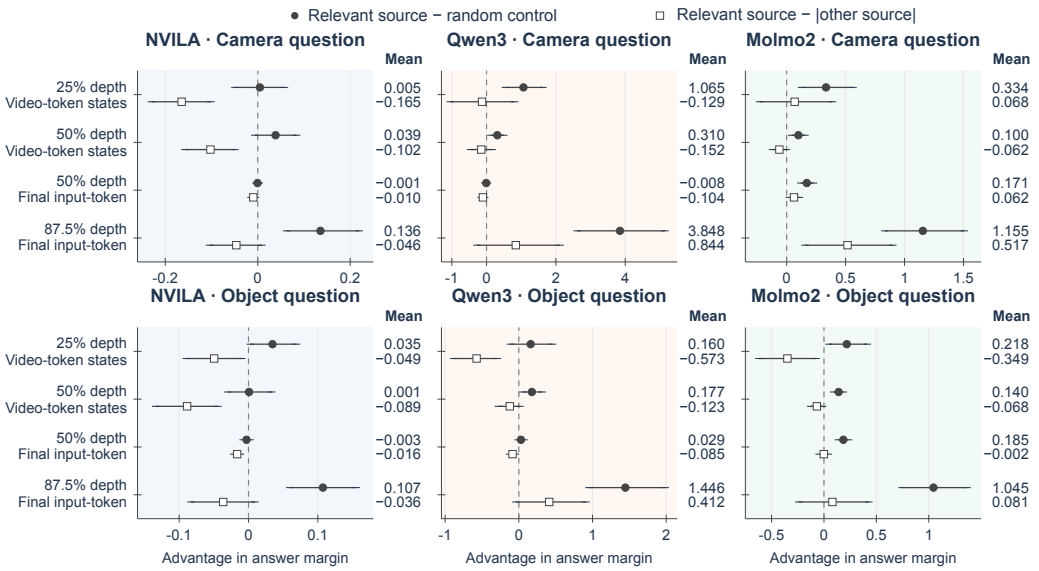


Figure A4: **Selectivity of additive hidden-state source interventions.** Vertical labels specify intervention depth and token scope. Filled circles and open squares show answer-margin advantages over norm-matched random and absolute other-source effects, respectively. Bars denote 95% bootstrap CIs; adjacent values are means. Both CIs must exceed zero for selectivity.

Molmo2 exhibits camera-selective effects at the final input-token position at 50% and 87.5% layer depth. The remaining tested camera/object combinations do not satisfy both selectivity criteria, and reliable object-selective effects are not established across model families. Thus, although source information can be partially recovered, additive interventions do not provide consistent evidence that it is used source-selectively by the native decoder. This result does not rule out nonlinear source representations or source-selective effects at other untested locations or layers.

B.5 Matched Motion Interactions and Interpretation Boundaries

B.5.1 Four-Condition Construction and Interaction Estimand

For each base scene, we construct four matched videos by varying camera and target motion while keeping the underlying scene construction fixed. We denote these conditions by V_{00} (neither), V_{10} (camera-only), V_{01} (object-only), and V_{11} (both). **Appendix** Fig. A5 presents the rendered inputs and target contours as experimental controls rather than model-performance results.

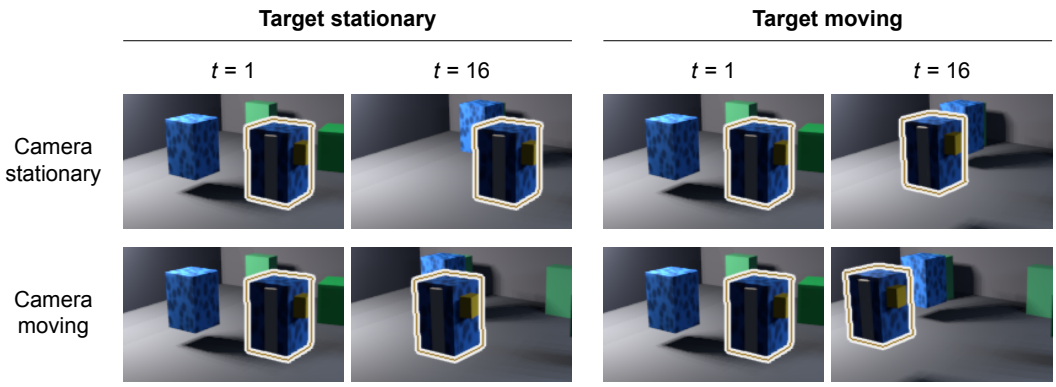


Figure A5: **Matched inputs for the interaction diagnostic.** Four variants of the same base scene isolate camera motion, target motion, their concurrence, and the no-motion control. Displayed frames and target contours are taken from the rendered data. Matching the scene construction enables within-unit comparisons of answer evidence across motion conditions.

For camera- and object-question answer margins m_c and m_o, respectively, we define standardized interactions that measure how each source changes the other source's incremental evidence:

$$\mathcal{I}_c = \frac{[m_c(V_{11}) - m_c(V_{01})] - [m_c(V_{10}) - m_c(V_{00})]}{\sigma_c},$$
$$\mathcal{I}_o = \frac{[m_o(V_{11}) - m_o(V_{10})] - [m_o(V_{01}) - m_o(V_{00})]}{\sigma_o}, \tag{A1}$$

where σ_c and σ_o are the corresponding model/task margin standard deviations over the four-condition observations. A negative interaction indicates that the evidence increment associated with one source is smaller when the other source is also moving. It is not an accuracy difference and does not, by itself, identify a particular neural mechanism in the evaluated model under this diagnostic.

B.5.2 Independent-Content Replication

Table A9: **Non-additive source interactions under concurrent motion.** Each row reports the standardized interaction for the corresponding source-question answer margin. All six intervals lie below zero in the first construction, and five of six do so in the independent construction. Only the independent NVILA object interval crosses zero; values are rounded to three decimal places.

Model	Source	First construction Effect [95% CI]	Independent construction Effect [95% CI]
NVILA	Camera	−0.593 [−0.924, −0.294]	−0.489 [−0.843, −0.145]
	Object	−0.340 [−0.542, −0.150]	−0.087 [−0.281, 0.108]
Qwen3	Camera	−0.625 [−0.861, −0.415]	−0.619 [−0.806, −0.435]
	Object	−0.382 [−0.663, −0.122]	−0.397 [−0.698, −0.108]
Molmo2	Camera	−0.282 [−0.530, −0.042]	−0.264 [−0.458, −0.080]
	Object	−0.891 [−1.219, −0.591]	−0.783 [−1.094, −0.468]

Both constructions contain 30 base units with four conditions each. The second construction is disjoint from the first in decoded visual content. Its answer-margin analysis was specified after the first result; it is an independent-content replication, not an untouched preregistered confirmation. The second construction also supports probe experiments, but those probes do not use the answer margins or questions analysed here. **Appendix** Tab. A9 reports estimates and intervals for both constructions under the same reporting convention used throughout this analysis.

B.5.3 Single-Source Competence Check

To assess whether interaction failures can be attributed specifically to combining already-mastered source judgments, we additionally audit single-source conditions on a development set. Three motion amplitudes (0.9, 1.25, and 1.55) are evaluated using 10 base units per amplitude. Camera-only and object-only performance uses exact state accuracy, which requires correctly identifying both sources. **Appendix** Fig. A6 shows the prespecified 70% competence threshold.

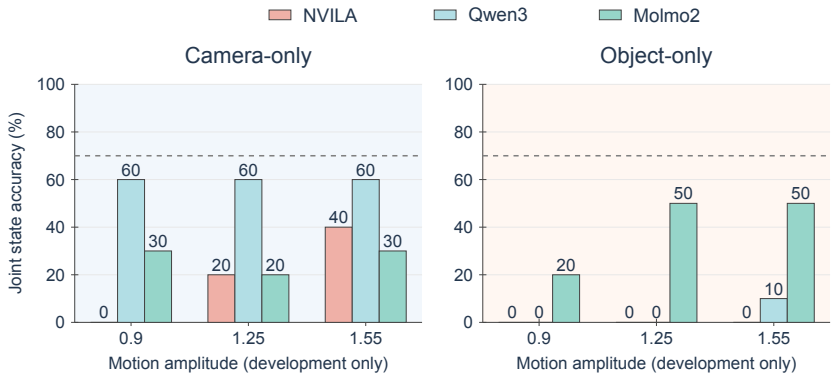


Figure A6: **Development-set check of single-source competence.** Left and right panels report exact state accuracy for camera-only and object-only conditions, respectively, across three motion amplitudes. Each amplitude uses 10 base units. The dashed line marks the prespecified 70% competence threshold; these are development diagnostics, not main-benchmark scores.

The tested configurations do not reach the prespecified competence threshold. Therefore, the negative interactions cannot be attributed specifically to a failure to compose otherwise mastered single-source judgments. The supported conclusion is narrower: source-specific answer evidence combines non-additively under concurrent motion in the tested matched conditions. This check constrains the interpretation of the interaction result without changing the measured interaction effects.

C Additional Implementation Details

C.1 Backbone Instantiations

We use the short model names established in the Diagnostic Setup throughout the main text and result tables. The 7B and 8B designations are nominal released-model scales rather than exact loaded-parameter counts; for reproducibility, **Appendix** Tab. A10 records each checkpoint and revision, loaded parameter count, and MoSDeR-optimized parameter count used in all experiments.

Table A10: **Backbone instantiations and parameter scales.** Parameter counts follow the model implementations and checkpoint revisions used in our experiments.

Short name	Exact checkpoint and revision	Nominal scale	Loaded parameters	MoSDeR trainable
NVILA	NVILA-Lite-8B `ea3c8b6d50a4`	8B	8,087,063,152	2,572,576 (0.0318%)
Qwen3	Qwen3-VL-8B-Instruct `0c351dd01ed8`	8B	8,767,123,696	2,982,176 (0.0340%)
Molmo2	Molmo2-O-7B `784410650d12`	7B	7,760,786,896	3,014,944 (0.0388%)

Loaded parameters count the original VLM backbone, excluding added MoSDeR modules. MoSDeR trainable counts the union of distinct parameters optimized across source factorization, shared adaptation, and residual refinement, rather than the parameters updated simultaneously in any single stage. Percentages are relative to the loaded backbone parameters. The separately fitted and subsequently frozen allocator is excluded: 458,817 parameters for NVILA and 524,353 for Qwen3 and Molmo2; checkpoint revisions use 12-character prefixes.

C.2 Module Instantiation

The implementation preserves each VLM's native visual tokenization and language decoding interfaces. The query-conditioned allocator projects frozen question embeddings and visual tokens into a shared matching space, combines normalized similarity with a learned visual bias, and predicts M_q. Sparse target-box occupancy provides balanced binary supervision, after which the allocator is frozen for every later stage.

For each source $s \in \{o, c\}$, the temporal source module concatenates normalized features with first-order temporal differences. The source-specific two-layer bottleneck MLP P_s uses GELU activation, after which T_s mixes the representation over time. Camera and object pathways use separate parameters but aligned temporal indexing and output dimensions for subsequent composition.

Each factor enters through complementary interfaces: the visual interface broadcasts the aligned factor to its allocated token region, while the language interface modulates a dedicated low-rank decoder branch. The bounded residual Δ_s is zero-initialized, so refinement starts from the preceding solution without an abrupt perturbation during the subsequent optimization stage.

C.3 Stagewise Parameter Updates

Training separates semantic grounding from later task adaptation. First, balanced occupancy supervision fits the query-conditioned allocator; the question representation remains frozen, and the fitted allocator is held fixed thereafter. Second, source-factor grounding activates only the pathway associated with the supervised source and optimizes its temporal adapter, conditional decoder branch, and visual injection using Eq. (6). Third, the grounded source components are frozen while the shared language branch is trained with the field-balanced objective in Eq. (8).

Residual refinement is deliberately isolated from the preceding stages. All previously learned parameters remain fixed, and only γ_s and R_s in Eq. (9) are optimized. During downstream adaptation, the temporal source modules, source-conditioned decoder factors and gates, shared language branch, and explicit residual modules are jointly updated using Eq. (11). The base VLM, fitted allocator, and factor intercepts remain frozen. Source-supervised rehearsal preserves the learned camera and object organization throughout downstream adaptation on the OmniVCHall benchmark.

C.4 Reproducibility and Comparison Protocol

Common Comparison Controls. Within each family, trained methods share the raw VLM and ordered MoSA split; all methods use the matched task information, answer space, and evaluator in **Appendix** Tab. A6.

Method-Faithful Reproductions. For 4DP-QA (Cho et al., 2026) and 4D-RGPT (Yang et al., 2026), we construct backbone-specific MoSA reproductions with separate runs for every backbone. Each uses the same split, target specification, answer space, and scoring as MoSDeR while retaining the method's defining architecture and training protocol; pretrained components retain their original roles rather than serving as transferred task checkpoints.

Downstream Baseline. On OmniVCHall, we follow public TriCD (Xing et al., 2026) and its perturbation design. All methods share inputs, decoding, and evaluation under the official protocol.

Reproducibility Scope. **Appendix** Tab. A10 lists backbones; **Appendices** C.2 and C.3 define modules and update scope. Shared task settings follow the common protocol, while every method retains backbone-native preprocessing and tokenization.

C.5 Inference Interface

For general video-language reasoning, inference uses only the original video and language query. The learned allocator replaces box-derived assignments, after which the temporal source modules, visual injection, and conditional decoder branches execute within a single forward generation path. No physical trajectories, source labels, or intermediate source states are required. For controlled MoSA evaluation, MoSDeR instantiates M_q from the common target-box track, which identifies the queried target without exposing trajectories, source labels, or answers.

D Additional Experimental Results

D.1 Within-Source Motion-State Analysis

To examine whether the MoSA results can be explained by cross-dataset motion-state correlations, we further evaluate Molmo2-based methods state-wise on the AV2 subset, where all four motion states occur within the same source dataset.

Table A11: **Within-source motion-state analysis on AV2 (MoSA) with Molmo2.** Overall pools all samples, while Macro averages the four motion states. Std and Min denote the standard deviation and minimum of state-wise accuracies, respectively.

Method	Neither	Camera	Object	Both	Overall	Macro	Std ↓	Min
Raw VLM	64.29	32.97	55.00	31.35	38.75	45.90	14.14	31.35
Fine-tuning	52.86	75.14	55.00	54.05	62.08	59.26	9.20	52.86
4DP-QA (Cho et al., 2026)	**68.57**	60.54	**92.50**	83.24	73.13	**76.21**	12.44	60.54
4D-RGPT (Yang et al., 2026)	47.14	**75.68**	55.00	51.89	60.62	57.43	10.90	47.14
MoSDeR (Ours)	**68.57**	74.05	62.50	**86.49**	**77.08**	72.90	**8.84**	**62.50**

As shown in **Appendix** Tab. A11, MoSDeR achieves the highest pooled accuracy (77.08%), outperforming 4DP-QA by 3.95 percentage points, and obtains the highest accuracy on concurrent motion

(*Both*: 86.49% versus 83.24%). It also yields the lowest dispersion across state-wise accuracies (Std: 8.84 points) and the highest worst-state accuracy (Min: 62.50%), indicating a more balanced performance profile across motion states.

The advantage is not uniform across all states: 4DP-QA attains a higher macro-average accuracy (76.21% versus 72.90%) based on 4DP-QA's macro edge is driven by *Object-only* performance, which is disproportionately higher *Object-only* accuracy (92.50% versus 62.50%). Thus, MoSDeR's advantage on AV2 arises from stronger pooled and concurrent-motion performance together with a higher performance floor and exhibits more balanced cross-state behavior. Because all samples in this analysis come from the same source dataset, cross-dataset source identity cannot account for these state-wise differences. The results therefore provide complementary evidence that MoSDeR's gains are not solely attributable to dataset-source priors.

D.2 Downstream Qualitative Analysis

The OmniVCHall examples in **Appendix** Fig. A7 illustrate corrected temporal answer selection after downstream adaptation without supplied target boxes. For the cat, MoSDeR preserves the shared scene description while capturing the change from facing the camera to facing sideways. For the pool player, it rejects both proposed event orders without asserting a replacement chronology. The displayed texts expand recorded option selections rather than represent free-form explanations. Together, these cases show complementary temporal corrections: recognizing an orientation change and rejecting unsupported event orderings in questions without explicit motion-source labels.

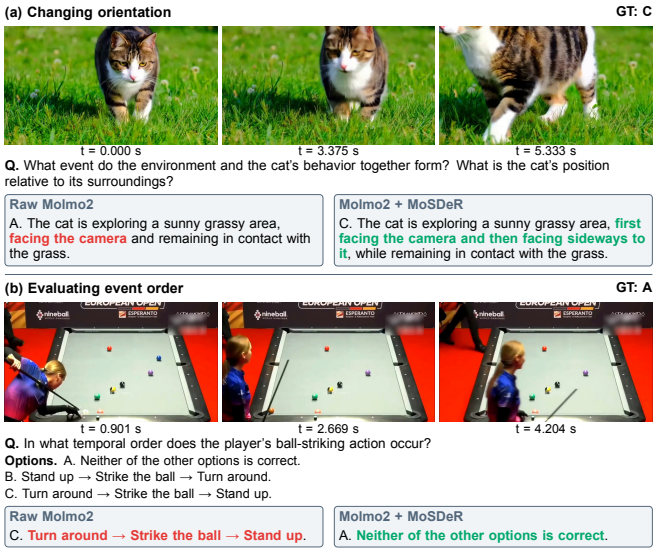


Figure A7: **Qualitative results on OmniVCHall.** MoSDeR corrects answers without supplied target boxes.

D.3 Computational Efficiency

Appendix Tab. A12 reports the complete parameter and runtime measurements corresponding to the main-text efficiency analysis. Across the three backbones, MoSDeR uses 2.573–3.015M trainable parameters, 4.2–12.5% fewer than budget-matched fine-tuning, while its latency differs by only -0.9% to $+6.6\%$. Counts exclude the frozen allocator (0.46–0.52M), which is fitted once and never updated during adaptation; the comparison therefore reflects adapted capacity. 4DP-QA provides lower latency but requires 35.6–44.2$\times$ more trainable parameters than MoSDeR; 4D-RGPT requires 8.7–9.0$\times$ more and is consistently slower. Peak memory is the main trade-off: MoSDeR uses 27.54–33.05 GB, exceeding budget-matched fine-tuning across all three backbones despite comparable latency and throughput across all measured backbone–method configurations.

Table A12: **Parameter and runtime efficiency on MoSA.** The table reports trainable parameters, peak memory, inference latency, and throughput for all three backbones.

Model / Method	Adaptation	Runtime		
	Trainable (M)	Peak Mem. (GB)	Latency (s)	Samples/s
NVILA + Fine-tuning	2.884	25.09	1.481	0.675
NVILA + 4DP-QA (Cho et al., 2026)	113.789	33.96	0.887	1.127
NVILA + 4D-RGPT (Yang et al., 2026)	23.074	34.69	2.465	0.406
NVILA + MoSDeR (Ours)	2.573	33.05	1.579	0.633
Qwen3 + Fine-tuning	3.408	22.08	3.298	0.303
Qwen3 + 4DP-QA (Cho et al., 2026)	127.159	28.87	0.774	1.292
Qwen3 + 4D-RGPT (Yang et al., 2026)	26.482	35.25	5.131	0.195
Qwen3 + MoSDeR (Ours)	2.982	30.43	3.267	0.306
Molmo2 + Fine-tuning	3.146	19.21	1.203	0.831
Molmo2 + 4DP-QA (Cho et al., 2026)	107.236	22.77	0.817	1.224
Molmo2 + 4D-RGPT (Yang et al., 2026)	26.220	33.11	2.161	0.463
Molmo2 + MoSDeR (Ours)	3.015	27.54	1.266	0.790

D.4 Hyperparameter Sensitivity

Appendix Tab. A13 shows a broad stable region rather than a single favorable setting. Overall accuracy is 88.92/89.17/87.08% for $d_b = 16/32/64$ and 88.83/89.17/87.75% for $r = 4/8/16$, indicating that moderate capacity is sufficient and larger settings provide no consistent benefit. The injection scale is especially stable: every value from 0.010 to 0.050 yields 89.17% Overall despite redistributing accuracy across states, while 0.075 remains at 88.75%. Object and Both accuracy are nearly invariant to η, consistent with the ablation finding that decoder routing is the primary pathway and visual injection provides a complementary refinement (see Tab. 3). The residual bound likewise forms a plateau from 0.025 to 0.150 (88.50–89.17%), with a clearer decline only at 0.200 (86.83%). State-wise changes are more pronounced than the aggregate trend: $d_b = 16$ favors Neither and Camera but weakens Both, whereas $r = 4$ improves Both while reducing Object. Performance remains stable without narrow tuning: the 12–32-frame settings stay within 3.34% of the selected result, confirming a stable sampling range.

Table A13: **Hyperparameter sensitivity on MoSA validation set.** Gray rows mark selected settings.

Value	Neither	Camera	Object	Both	Overall
Bottleneck width d_b					
16	94.33	95.00	93.33	73.00	88.92
32	**87.00**	**83.67**	**96.00**	**90.00**	**89.17**
64	85.33	82.67	93.67	86.67	87.08
Decoder rank r					
4	86.33	83.33	93.67	92.00	88.83
8	**87.00**	**83.67**	**96.00**	**90.00**	**89.17**
16	85.67	83.33	94.67	87.33	87.75
Injection scale η					
0.010	83.33	87.33	96.00	90.00	89.17
0.025	84.67	86.00	96.00	90.00	89.17
0.050	**87.00**	**83.67**	**96.00**	**90.00**	**89.17**
0.075	86.33	82.33	96.00	90.33	88.75
0.100	85.67	80.00	95.67	89.33	87.67
Sampled frames					
8	84.33	85.67	95.33	87.00	88.08
12	85.33	85.67	95.67	87.67	88.58
20	**87.00**	**83.67**	**96.00**	**90.00**	**89.17**
32	88.33	75.67	96.00	83.33	85.83
Residual bound β					
0.025	86.33	85.33	93.67	88.67	88.50
0.050	86.67	85.00	94.67	89.00	88.83
0.100	**87.00**	**83.67**	**96.00**	**90.00**	**89.17**
0.150	85.33	81.33	96.33	91.33	88.58
0.200	81.33	79.67	95.33	91.00	86.83

D.5 Qualitative Case Audit on MoSA

Appendix Tab. A14 audits fixed counts. Relative to Raw Molmo2, MoSDeR raises Camera-only/Object-only/Both-source correct cases from 10/65/3 to 73/92/89. The Both-source gain supports non-exclusive composition over single-source selection. Target-switch pairs improve from 12 to 52 but remain lowest, indicating stronger yet target-sensitive binding. This balanced audit complements full-distribution evaluation and localizes remaining attribution errors.

Table A14: **Audit of representative qualitative cases.** Fixed Molmo2 validation audit counts under the common evaluation protocol in **Appendix** Tab. A6.

Case type	Evaluated	Raw correct	MoSDeR correct	Primary question
Camera-only motion	100	10/100	73/100	Is viewpoint change camera-induced?
Object-only motion	100	65/100	92/100	Is target motion separated from static view?
Both-source motion	100	3/100	89/100	Are both sources identified?
Target-switch pairs	100	12/100	52/100	Does judgment change with target?