TCMU: TEMPORAL CAMERA MOTION UNDERSTANDING

Anonymous authors
Paper under double-blind review

ABSTRACT

Camera motion is temporal and compositional: multiple operations can overlap, change speed, and transition within a single shot. Yet shot-level labels leave this structure unspecified. We introduce Temporal Camera Motion Understanding (TCMU), which predicts a sequence of temporal segments describing the types, directions, and speeds of all active camera motions. To support learning and evaluation, we construct a training dataset of approximately 292K human-annotated video clips and TCMU-Bench, a separate benchmark of 1,000 shots with 2,590 manually annotated segments. We develop two approaches. **TCMU-Geo** derives motion timelines from reconstructed geometry by separating depth-independent and depth-dependent components of camera-induced image motion. **TCMU-VLM** learns to generate these timelines directly from video using the annotated training dataset. On TCMU-Bench, Geo achieves 60.7 frame F1 versus 32.8 for the pose-based SpatialVID baseline. The VLM achieves 71.5 frame F1 versus 25.4 for the strongest evaluated zero-shot VLM, alongside 70.0 Event-F1 at temporal IoU 0.5 and 84.9 Boundary-F1 at 0.5s tolerance. The VLM is stronger overall in recognition and temporal localization, while condition and concurrency analyses show Geo's advantages in parallax-based motion discrimination and directed-motion recall under dense concurrency. Paired event analysis shows that Geo recovers directed-motion events missed by the VLM, motivating geometry-assisted human review of training annotations. Training on the reviewed annotations improves VLM recognition and event localization but reduces boundary-localization accuracy. We will release TCMU-Bench, the evaluation code, the distilled TCMU-VLM-2B checkpoint, and the complete TCMU-Geo code.

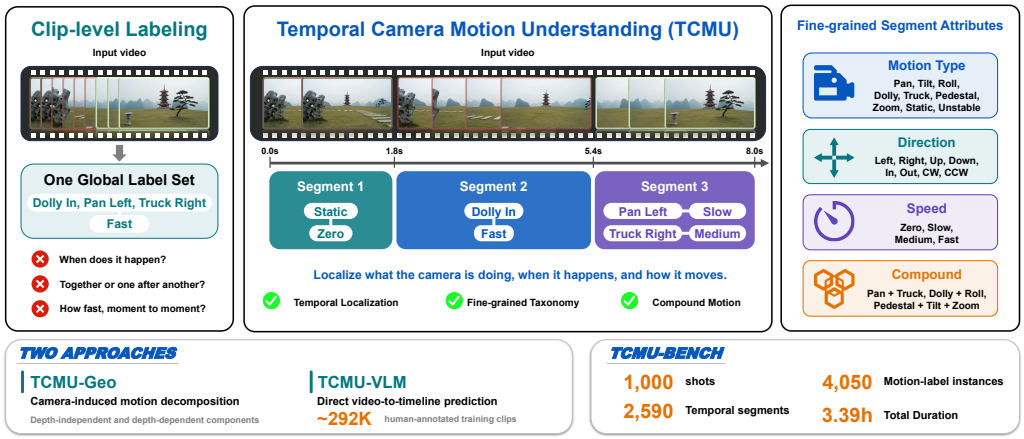


Figure 1: **Task, data, and methods.** TCMU represents a shot as a sequence of intervals with motion types, directions, and speeds, including simultaneous operations. TCMU-Geo derives this timeline from reconstructed geometry; TCMU-VLM learns it from human annotations. Both are evaluated on the separate TCMU-Bench. The lower panels summarize the two approaches, the approximately 292K human-annotated training clips, and the scale of TCMU-Bench.

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.

1 INTRODUCTION

Camera motion unfolds over time, even within a single uninterrupted shot. A camera may remain still, dolly forward, and then pan while continuing to translate. A clip-level label set can identify all these motions yet leave their temporal organization unresolved: the same set describes motions occurring simultaneously or in sequence, and does not associate each speed with its interval. Recent benchmarks improve semantic recognition through fine-grained annotations and VLM training (Lin et al., 2026; Tan et al., 2026), while fixed-grid approaches provide temporally sampled descriptions (Feng et al., 2026). Fine-grained camera-motion descriptions support video-text retrieval (Lin et al., 2026) and language-guided camera control and video generation (Liu et al., 2024; Yang et al., 2026; Zhao et al., 2026).

We introduce **Temporal Camera Motion Understanding (TCMU)**: given a single shot, predict an ordered sequence of intervals, each containing the active camera motions and their type, direction, and speed. As illustrated in Figure 1, this representation captures both sequential changes and within-interval composition. It requires recognizing concurrent components and localizing their temporal extent.

We construct approximately 292K **human-annotated video clips** to learn this structured prediction task. For evaluation, we build **TCMU-Bench**: 1,000 shots with 2,590 manually annotated segments, covering nine motion types and independently checked by three reviewers. The benchmark evaluates frame-level recognition, motion-event localization, and state-boundary localization, with subsets organized by motion structure and challenging visual conditions.

We investigate two approaches. **TCMU-Geo** derives motion timelines from reconstructed geometry by separating depth-independent and depth-dependent image motion, using a pretrained geometry estimator (Wang et al., 2026b) without task-specific recognition training. **TCMU-VLM** learns to generate the same timelines through supervised fine-tuning on the human-annotated training dataset.

Both methods outperform their respective baselines in all three metrics. The VLM is stronger overall, while Geo leads in parallax-based frame recognition and directed-motion recall under dense concurrency. As an additional application, geometry-assisted annotation review improves VLM recognition and event localization, with a decline in boundary accuracy.

Our contributions are:

- **Task and data.** We define TCMU, construct a human-annotated dataset for model training, and introduce TCMU-Bench for evaluating recognition and temporal localization.
- **Two methods.** We develop TCMU-Geo, which decodes motion timelines by separating depth-independent and depth-dependent camera-induced image motion, and TCMU-VLM, which learns timeline prediction from human annotations.

We will release TCMU-Bench, the evaluation code, the distilled TCMU-VLM-2B checkpoint, and the complete TCMU-Geo code.

2 RELATED WORK

Camera-motion representations. Early methods classify temporally smoothed image-motion signals into coarse operations (Guironnet et al., 2006; Duan et al., 2006), while cinematography-oriented work develops motion taxonomies and shot-level recognition datasets (Wang & Cheong, 2009; Hasan et al., 2014; Petrogianni et al., 2022). MovieNet (Huang et al., 2020) treats motion as clip-level attributes; CameraBench (Lin et al., 2026) provides fine-grained expert labels and annotation-trained VLMs; and ACaM studies atomic natural-language camera movements (Tan et al., 2026). These approaches do not jointly represent arbitrary event intervals, concurrent operations, and interval-specific speed. Trajectory-oriented work such as GenDoP and SpatialVID represents physical camera paths, with SpatialVID converting pose increments into per-step instructions (Zhang et al., 2025; Wang et al., 2026a); geometry-guided VLMs also predict motion on a fixed temporal grid (Feng et al., 2026).

Temporal prediction and geometric evidence. Temporal action localization predicts event intervals (Zhang et al., 2022; Shi et al., 2023), dense video captioning associates intervals with descrip-

tions (Yang et al., 2023), and autoregressive VLMs generate temporally grounded content (Guo et al., 2024; Wang et al., 2026c). TCMU requires a complete temporal partition of a shot, with concurrent motion types, directions, and interval-specific speeds specified within each interval. Classical motion-field analysis relates image velocities to camera motion and scene structure (Longuet-Higgins & Prazdny, 1980), while parametric flow models support qualitative camera-motion readout (Bouthemy et al., 1999; Park et al., 2004). SfM (Schönberger & Frahm, 2016) and learned estimators such as MegaSaM and VGGT-Ω recover cameras and scene structure (Li et al., 2025; Wang et al., 2026b); geometry-aware VLMs incorporate such evidence through prompts or learned representations (Feng et al., 2026; Fan et al., 2026).

3 Task: Temporal Camera Motion Understanding

Given a single continuous shot V of duration T, TCMU predicts a sequence of motion segments:

$$\mathcal{Y}(V) = \left\{(t_i^{\text{start}}, t_i^{\text{end}}, \mathcal{B}_i)\right\}_{i=1}^{N}.$$

The intervals partition the shot: $t_1^{\text{start}} = 0$, $t_N^{\text{end}} = T$, and $t_i^{\text{end}} = t_{i+1}^{\text{start}}$. A new segment begins whenever an active motion type, direction, or speed changes. Boundaries may occur at arbitrary timestamps rather than on a fixed temporal grid. Videos containing cuts are split into continuous shots before prediction.

Vocabulary. The motion names follow established camera-motion taxonomies (Petrogianni et al., 2022; Lin et al., 2026; Tan et al., 2026), with operational definitions adapted to TCMU. The closed vocabulary contains nine types: `Pan`, `Tilt`, and `Roll` describe rotation; `Truck`, `Pedestal`, and `Dolly` describe translation; `Zoom` describes focal-length change; `Static` denotes a hold without perceptible shake; and `Unstable` denotes irregular shake without a steady direction. Directions use the camera's local coordinate system. `Static` has speed `zero` only. All other types take `slow`, `medium`, or `fast`, corresponding to subtle but confirmable, clearly perceptible, or rapid motion. Perceptible irregular shake is labelled `Unstable` even at low magnitude, so the `Static`–`Unstable` distinction depends on perceptibility.

Concurrent motions. Each $\mathcal{B}_i$ is a set of $(\text{type}, \text{direction}, \text{speed})$ tuples drawn from this vocabulary, describing all camera operations active throughout the interval. For example, `Truck left` and `Pan right` can coexist in one segment. Each motion type may appear at most once in $\mathcal{B}_i$, with a single direction and speed; for example, `Pan left` and `Pan right` cannot coexist within the same segment. `Static` and `Unstable` each occupy a segment alone and cannot co-occur with other types. The vocabulary and its parallel and sequential composition are illustrated in Appendix A.3 (Figure 5).

4 Data and Benchmark

4.1 Human-annotated training dataset

We construct approximately 292K distinct human-annotated clips from curated film and television footage, each containing one continuous shot and a complete timeline of motion types, directions, and speeds. Initial quality control uses 10% random sampling; annotation and review procedures are detailed in Appendix A.1. This dataset is used to train TCMU-VLM. Training data and TCMU-Bench are disjoint at the source-video level (Appendix A.2).

4.2 TCMU-Bench

TCMU-Bench contains $1{,}000$ continuous shots spanning 3.39 hours and $2{,}590$ annotated segments. Clips are drawn from GenDoP (Zhang et al., 2025) and DynPose-100K (Rockwell et al., 2025), supplemented with 23 videos generated using Seedance 2.0 to cover underrepresented conditions, including dolly zooms, picture-in-picture, and heavy visual effects. All generated clips belong to the Challenging subset. We release source-video references and annotations for the existing datasets, and the generated evaluation videos, with the output schema provided in Appendix B.3.

Benchmark annotation and quality control. Three annotators with film or photography backgrounds label the shots using the vocabulary in Sec. 3, and three reviewers independently inspect each

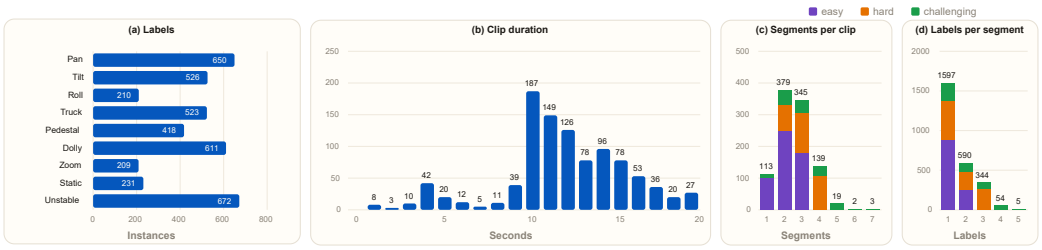


Figure 2: **Composition of TCMU-Bench** (1,000 shots and 2,590 temporal segments). (a) Number of annotated instances for each camera-motion type, pooling directions. (b) Distribution of shot duration. (c) Number of temporal segments per shot. (d) Number of simultaneously active motion labels per segment. Colors in (c–d) denote the *easy*, *hard*, and *challenging* subsets.

shot. Segmentation, revision, and disagreement-resolution procedures are detailed in Appendix A.4. Shots with a tied vote on any disputed label are excluded from the final benchmark.

Scale and composition. The benchmark contains 4,050 motion-label instances across all nine types; 993 of its 2,590 segments contain multiple directed operations. Figure 2 summarizes label, duration, segmentation, and concurrency distributions, with detailed combination coverage in Appendix A.5.

Benchmark subsets. We organize the benchmark into three disjoint subsets based on motion structure and selected challenging conditions. Easy and Hard distinguish motion-structure complexity; Challenging groups clips selected for specific visual or structural conditions.

Challenging contains 153 shots selected during curation, each assigned one condition from three families: *parallax discriminations*, *reconstruction-challenging conditions*, and *structural complexity*. These families organize the condition-level analysis in Sec. 7.2; their detailed composition is provided in Appendix A.4 (Table 5).

The remaining shots are divided by motion structure. *Easy* contains 526 shots with at most three segments. Excluding `Static` and `Unstable`, each shot involves either a single motion type or a compound of two motion types. *Hard* contains the remaining 321 shots with more segments, three or more concurrent motion types, or transitions between distinct motion configurations.

We evaluate complete shots in each subset under the same recognition and temporal-localization protocol (Sec. 7.2). The subsets also differ in their proportions of `Static` and `Unstable` (Appendix A.4).

5 Two Solutions to TCMU

TCMU-Geo derives motion labels from reconstructed camera and scene geometry, whereas TCMU-VLM learns to predict the same labels directly from annotated videos. Both output the segment sequence defined in Sec. 3 and operate independently at inference time. Their common output space allows us to compare recognition, event and boundary localization, and recovery of concurrent motions across the two approaches.

5.1 TCMU-Geo: Camera-induced motion fields for structured prediction

TCMU-Geo uses a pretrained geometry estimator followed by a deterministic readout, without task-specific recognition training. We construct a *camera-induced motion field* by reprojecting reconstructed scene points under the estimated camera transformation and intrinsics while holding the points fixed in the scene. This field records image-plane displacements between adjacent frames and is computed directly from geometry, without a separate optical-flow estimator. The four stages are geometry recovery, motion-field synthesis, channel decomposition, and temporal decoding (Figure 3).

From geometry to image motion. Directly thresholding reconstructed pose increments presents two difficulties. First, monocular translation is recovered only up to an unknown scale (Bian et al., 2019): the same physical camera motion may have different reconstructed magnitudes across videos,

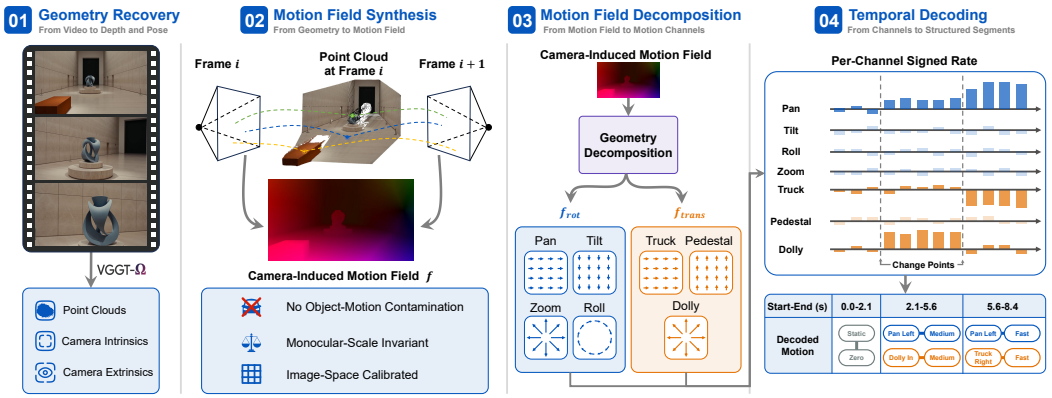


Figure 3: **TCMU-Geo.** The pipeline recovers scene geometry and synthesizes the full camera-induced motion field and its depth-independent component by reprojection. It projects the depth-independent component and the depth-dependent residual onto motion bases, then decodes the resulting channel rates into temporal segments.

making a common pose threshold unreliable. Second, translation-induced image motion depends on scene depth, whereas rotational image motion is depth-independent (Longuet-Higgins & Prazdny, 1980). For example, the same lateral truck can move a nearby subject substantially across the frame while producing barely perceptible displacement in a distant landscape. We therefore measure the image motion induced by the estimated camera. Appendix B.1 (Figure 6) illustrates this distinction.

VGGT-Ω (Wang et al., 2026b) predicts depth, intrinsics, and camera poses. For adjacent frames with depth Z_t, intrinsics K_t, K_{t+1}, and relative pose $(R, \mathbf{t})$, we synthesize

$$\mathbf{f}(\mathbf{x}) = \pi\big(K_{t+1}\left(RZ_t(\mathbf{x})K_t^{-1}\tilde{\mathbf{x}} + \mathbf{t}\right)\big) - \mathbf{x}, \qquad \mathbf{f}_{\mathrm{rot}}(\mathbf{x}) = \pi\big(K_{t+1}RK_t^{-1}\tilde{\mathbf{x}}\big) - \mathbf{x}, \tag{1}$$

where $\mathbf{x} = (u, v)^\top$ is a pixel coordinate, $\tilde{\mathbf{x}} = (u, v, 1)^\top$ is its homogeneous form, and π denotes perspective division. The second field contains the effects of rotation and changing intrinsics and is depth-independent. The residual $\mathbf{f}_{\mathrm{trans}} = \mathbf{f} - \mathbf{f}_{\mathrm{rot}}$ captures the depth-dependent contribution of translation.

Jointly rescaling depth and translation leaves the synthesized field unchanged, removing their shared monocular scale ambiguity. Because the field describes the projected displacement of scene points held fixed under the estimated camera motion, it does not measure their observed motion and requires no separate object-motion segmenter. Errors in reconstructed cameras or depth can nevertheless propagate to the motion labels.

From motion fields to motion channels. Following classical parametric camera-motion readouts (Bouthemy et al., 1999; Park et al., 2004), we fit each component by least squares using horizontal, vertical, radial, and tangential basis fields:

$$\mathbf{f}_n(x, y) \approx a\,\mathbf{e}_x + b\,\mathbf{e}_y + c[x, y]^\top + d[-y, x]^\top. \tag{2}$$

Coordinates and displacements are normalized by frame width, and rates are expressed per second. The coefficients (a, b, d, c) of the depth-independent stream provide `Pan`, `Tilt`, `Roll`, and `Zoom` channels; (a, b, c) of the translation stream provide `Truck`, `Pedestal`, and `Dolly` channels. Directions are assigned using the corresponding pose or focal-length changes.

This decomposition provides an explicit readout for visually similar operations arising from different geometric sources. Median motion-field magnitude and the translation-stream fit residual additionally support `Static` and `Unstable` decisions. The complete channel mapping is provided in Appendix B.1 (Table 8); inference settings and decision thresholds are specified in Appendix B.4 (Table 9).

Temporal decoding. Change-point detection (Killick et al., 2012; Truong et al., 2020) operates on the seven signed channel rates. Within each segment, a channel is scored by its mean signed rate,

allowing opposing fluctuations to cancel. All motion channels exceeding their respective thresholds are retained. Speed tiers use the same normalized amplitudes. The overall motion-field magnitude and fit residual determine the fallback state when no directed operation is selected.

5.2 TCMU-VLM: Learning from camera-motion annotations

Structured prediction. We fine-tune Qwen3-VL-8B-Instruct (Bai et al., 2025) to map video directly to the segment sequence of Sec. 3. The prompt specifies the closed vocabulary, direction and speed rules, and concurrent-motion constraints; the model outputs JSON following Appendix B.3. Input settings and training parameters are provided in Appendix B.2.

Training. We fine-tune the vision encoder, vision-language aligner, and language model on the human-annotated training dataset in Sec. 4.1. Joint training learns the mapping from video to timestamps and concurrent motion labels; the vision-side ablation is reported in Appendix D.10. Unless explicitly marked otherwise, *TCMU-VLM* denotes this 8B model trained on the original human annotations, which is used in comparisons with Geo. The reviewed-annotation variant is introduced in Sec. 7.4.

For release, we distill a 2B student from the reviewed 8B teacher using sequence-level knowledge distillation (Kim & Rush, 2016). Appendix B.2.1 specifies its separate teacher-generated training data.

5.3 Comparing geometric and learned representations

TCMU-Geo separates depth-independent and depth-dependent image motion, explicitly encoding the geometric distinction among rotation, translation, and focal-length change. TCMU-VLM learns the label mapping from appearance and temporal context without this decomposition. Comparing their predictions lets us examine how these different representations affect recognition and concurrent-motion recovery. Sections 7.2 and 7.3 analyze performance across visual conditions and motion composition.

6 Evaluation Protocol

We evaluate frame-level recognition, motion-event localization, and state-boundary localization. A shot with constant-speed `Pan Left` on $[0, 6)$ and concurrent `Truck Right` on $[2, 4)$ contains two motion events and state boundaries at 2s and 4s; both labels should be recognized at 3s. We pool TP/FP/FN across videos to compute micro precision, recall, and F1. The detailed evaluation protocol is provided in Appendix C.1.

Frame-level recognition. At $t = 0, 0.1, \ldots < T$, we compare predicted and annotated type–direction label sets over the full shot: 14 directed labels plus `Static` and `Unstable`. Shared, extra, and missing labels count as TP, FP, and FN. We additionally report direction-agnostic type recognition and type–direction recognition grouped by GT concurrency or annotated speed. Strict-speed evaluation requires complete type–direction–speed tuples to match.

Motion-event localization. On both GT and prediction sides, adjacent intervals with the same type–direction label are merged into maximal continuous events, ignoring speed and changes in concurrent labels. Within each video and label, we maximize one-to-one event matches satisfying $\mathrm{tIoU}(g, p) = |g \cap p|/|g \cup p| \geq \theta$. Matched events count as TP, unmatched predictions as FP, and unmatched GT events as FN. Event-F1 uses $\theta \in \{0.3, 0.5, 0.7\}$.

State-boundary localization. An internal boundary marks a change in the complete type–direction–speed state, including speed-only changes; shot endpoints are excluded. We maximize one-to-one matches between predicted and annotated boundaries within absolute error δ. Matched, extra, and missed boundaries count as TP, FP, and FN. Boundary-F1 uses $\delta = 0.5$s, with 0.25s and 1.0s as additional tolerances. Matches compare timestamps without requiring agreement of adjacent motion labels.

7 EXPERIMENTS

We evaluate eleven system configurations on the 1,000 shots of TCMU-Bench using the protocol in Sec. 6. Unless otherwise stated, the analyses below use TCMU-VLM trained on the original annotations. All VLMs use the same system prompt.

7.1 OVERALL RECOGNITION AND TEMPORAL LOCALIZATION

Compared systems. We compare against the pose-based SpatialVID (Wang et al., 2026a) and four zero-shot VLMs: Gemini-3-flash, Qwen3-VL-8B (Bai et al., 2025), InternVL3.5-8B, and LLaVA-Video-7B. SpatialVID converts MegaSaM poses into motion instructions and supports seven motion types, excluding `Zoom` and `Unstable`; our default Geo uses VGGT-Ω reconstruction. VLM input settings are provided in Appendix B.2.

Table 1: Recognition and temporal-localization results on TCMU-Bench (%). Frame P/R/F1 use type–direction labels at 0.1s sampling; Event-F1 uses the listed tIoU thresholds, and Boundary-F1 uses 0.5s tolerance. All methods are scored over the full nine-type vocabulary. Full event and boundary P/R/F1 are in Appendix D.1.

System	Frame P	R	F1	Event@0.3	@0.5	@0.7	Boundary@0.5s
Geometric methods							
SpatialVID	37.5	29.2	32.8	32.1	21.8	10.2	18.3
TCMU-Geo	56.4	65.8	60.7	55.0	48.9	37.2	32.5
Zero-shot VLMs							
Gemini-3-flash (zero-shot)	27.6	23.5	25.4	24.4	16.9	9.8	16.7
Qwen3-VL-8B (zero-shot)	27.7	18.8	22.4	19.1	15.3	9.8	3.3
InternVL3.5-8B (zero-shot)	11.0	6.8	8.4	7.4	4.7	2.6	9.1
LLaVA-Video-7B (zero-shot)	9.0	5.1	6.5	4.7	2.2	1.2	11.7
Human-supervised timeline prediction							
TCMU-VLM-8B (original)	81.2	63.8	71.5	71.9	70.0	66.1	84.9
Additional models: annotation review and distillation							
TCMU-VLM-8B (reviewed)	84.3	68.2	75.4	74.9	72.2	67.5	80.9
TCMU-VLM-2B	77.3	59.2	67.0	67.0	64.5	59.6	78.0

Frame recognition. Both methods improve over their respective baselines (Table 1). Geo's advantage persists with identical MegaSaM reconstruction outputs and the shared seven-type vocabulary (Appendix D.2, Table 17), supporting the value of its motion-field readout and decoding beyond reconstruction choice and label coverage. The VLM's improvement over its zero-shot backbone demonstrates the effectiveness of task-specific supervision. Relative to Geo, its higher precision but slightly lower recall suggests more selective motion-label prediction.

Temporal localization and speed. Both methods also improve event and boundary localization over their baselines. The VLM's Event-F1 falls less than Geo's as tIoU becomes stricter, and it leads in boundary localization, indicating stronger performance under stricter timing requirements. Requiring exact speed matches lowers both methods' frame F1, showing that recognizing an operation does not ensure accurate speed assignment (Appendix D.1).

External benchmark evaluation. On CameraBench's 15-category suite over 969 clips (Lin et al., 2026), the reviewed 8B checkpoint (Sec. 7.4) outperforms CameraBench-SFT in macro F1 after collapsing timelines to clip-level labels. This provides an external check of motion recognition; the category mapping and F1 protocol are detailed in Appendix D.7.

7.2 PERFORMANCE ACROSS BENCHMARK SUBSETS

Table 2 compares complete shots across Easy, Hard, and Challenging, with Challenging further divided into three condition families.

Table 2: Full-shot F1 scores across benchmark subsets (%). Frame uses type–direction labels at 0.1s sampling; Event uses tIoU 0.5, and Boundary uses 0.5s tolerance. The final three rows partition Challenging. The VLM uses the original annotations. Additional thresholds and tolerances are reported in Appendix D.4.

Subset	TCMU-Geo			TCMU-VLM		
	Frame	Event	Boundary	Frame	Event	Boundary
Easy	57.0	46.1	29.0	79.0	79.7	93.2
Hard	63.8	50.6	33.7	68.8	65.2	81.7
Challenging	62.6	51.3	36.8	59.5	58.7	74.0
Challenging conditions						
Parallax discriminations	60.8	52.2	36.6	51.8	59.2	77.9
Reconstruction-challenging	55.8	41.7	25.0	63.7	64.2	60.8
Structural complexity	67.4	55.1	42.2	62.2	56.4	77.9

Geo leads in frame recognition on parallax discriminations, consistent with its explicit separation of depth-independent and depth-dependent motion. The VLM leads under reconstruction-challenging conditions, consistent with Geo's dependence on reliable camera and scene estimates. Aggregate subset rankings also reflect label composition: removing `Static` and `Unstable` narrows the recognition gap on Easy (Appendix D.3, Table 18).

Geo also has higher frame F1 under structural complexity, yet the VLM leads in event and boundary F1 across every subset and condition family. Recovering motion components therefore does not by itself ensure accurate event extent or state-change timing. Temporal localization remains a limitation of Geo despite its recognition advantages in these conditions. This divergence supports evaluating component recognition and temporal localization separately.

7.3 COMPLEMENTARY ERRORS AND CONCURRENT-MOTION RECOVERY

Complementary event recovery. In the type view, Geo recovers more directed-motion events, while the VLM is stronger on `Static` and `Unstable` (Table 3). Exclusive matches follow the same split: each method recovers events missed by the other, with distinct strengths in directed operations and hold states. Per-type recovery and confusion results are provided in Appendices D.5 and D.6.

Table 3: Paired GT-event recovery at tIoU 0.5 in the type view. Directions are collapsed before event extraction. Holds comprise Static and Unstable. Each GT event belongs to one of the four outcome columns. Per-type results are in Appendix D.5.

Group	GT events	Geo R	VLM R	Both	Neither	Only Geo	Only VLM
All	3,249	50.9	63.1	1,062	609	591	987
Directed	2,346	66.0	56.5	989	461	560	336
Holds	903	11.5	80.2	73	148	31	651

Concurrency-dependent behavior. Geo has higher frame F1 at GT concurrency two through five, while the VLM leads at concurrency one. As concurrency increases, the VLM retains high precision but loses recall, indicating incomplete recovery of concurrent components. Geo's comparatively stable recall is consistent with retaining all motion channels exceeding their respective thresholds (Appendix D.9, Table 25). These observations concern the combinations represented in the benchmark; the highest concurrency level contains only five annotated segments (Appendix A.5).

Qualitative comparison. Figure 4b illustrates unsupported translations from Geo when the subject moves with the camera, while Figure 4c shows the VLM substituting `Truck` for `Pan`. These examples illustrate distinct failure modes: extra geometric components and confusion between visually similar operations (further examples in Appendices E.1 and E.2).

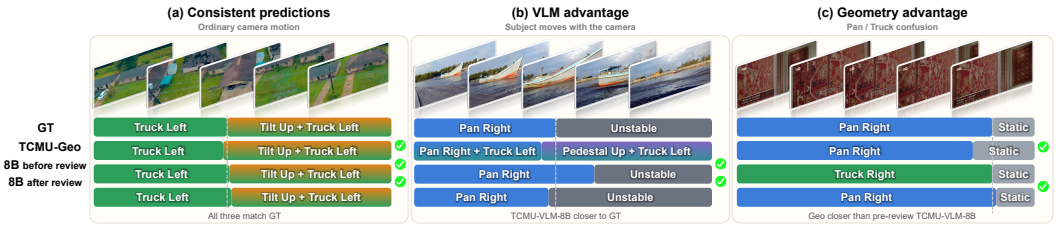


Figure 4: **Qualitative comparison of TCMU-Geo and TCMU-VLM.** Rows show GT, Geo, and the 8B VLM trained on original and reviewed annotations. The reviewed model is included as an additional reference. (a) All three models recover the displayed motion-label sequence. (b–c) show complementary recognition outcomes. Check marks indicate motion-label agreement; speed is omitted.

7.4 Application to Annotation Review

As an application, we use Geo to propose motion-label corrections for human review on approximately 2.3% of the training videos (Appendix A.6). Only human-confirmed directed-motion labels are revised, with speeds for newly added motions assigned by human reviewers. Existing speed annotations, hold states, temporal boundaries, training clips, and the 8B training configuration remain unchanged. Training on the reviewed annotations improves frame recognition and event localization but reduces boundary accuracy (Table 4). We further distill a 2B student from the reviewed teacher and will release it (Appendix B.2.1; Table 1).

Table 4: Annotation-review results (%). Frame uses type–direction labels at 0.1s sampling; event and boundary settings are tIoU 0.5 and 0.5s tolerance. Deltas are computed before rounding.

Metric	Original P	R	F1	Reviewed P	R	F1	ΔF1
Frame	81.2	63.8	71.5	84.3	68.2	75.4	4.0
Event@0.5	79.6	62.5	70.0	80.8	65.3	72.2	2.2
Boundary@0.5s	92.4	78.6	84.9	87.1	75.5	80.9	-4.0

8 Conclusion

We introduce TCMU to describe sequential and concurrent camera motions through a structured timeline of motion types, directions, and speeds. We construct approximately 292K human-annotated clips for training and introduce the independent TCMU-Bench, with 1,000 shots and 2,590 annotated segments for evaluating recognition, event localization, and state-change timing. TCMU-Geo derives timelines by separating depth-independent and depth-dependent camera-induced motion, while TCMU-VLM learns structured timeline prediction directly from video.

Both methods outperform their respective baselines. TCMU-VLM achieves 71.5 frame F1, 70.0 Event-F1 at tIoU 0.5, and 84.9 Boundary-F1 at 0.5s tolerance. Geo has advantages in parallax-based frame recognition and directed-motion recall under dense concurrency. As an additional application, geometry-assisted annotation review improves VLM recognition and event localization but reduces boundary accuracy. We will release TCMU-Bench, the evaluation code, the distilled TCMU-VLM-2B checkpoint, and the complete TCMU-Geo code. Limitations and future work are discussed in Appendix F.

AI Use Statement

The evaluation code was generated by AI. AI assistance was also used for manuscript editing and consistency checks between computed results and the manuscript. We used Seedance 2.0 to generate 23 evaluation videos for TCMU-Bench to cover underrepresented conditions. These videos were manually annotated and reviewed under the benchmark protocol and were not used for model training

or distillation. The authors are responsible for validating the evaluation protocol, code, annotations, results, and final text.

ETHICS STATEMENT

The human-annotated training dataset uses curated film and television footage. TCMU-Bench includes videos referenced by publicly available links from GenDoP and DynPose-100K, together with 23 videos generated using Seedance 2.0 to cover underrepresented evaluation conditions. For benchmark videos drawn from existing datasets, we provide original links and annotations rather than redistributing the source videos. The generated videos are used only for benchmark evaluation and are not used to train or distill the models evaluated in this work.

Annotators and reviewers were compensated for their work. The annotations concern camera motion only and do not identify individuals or infer personal or sensitive attributes. We did not conduct a dedicated privacy or sensitive-content screening of videos referenced from existing public datasets. For the generated videos distributed with the benchmark, we will provide a contact channel through which removal requests can be submitted.

REPRODUCIBILITY STATEMENT

Upon publication, we will release TCMU-Bench with source-video references, JSON annotations, and generated evaluation videos; the evaluation code; the distilled TCMU-VLM-2B checkpoint; and the complete TCMU-Geo implementation and configuration. The output schema is specified in Appendix B.3, model settings in Appendices B.2 and B.4, and the detailed evaluation protocol in Appendix C.1.

REFERENCES

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, 2025. URL https://arxiv.org/abs/2511.21631.

Jiawang Bian, Zhichao Li, Naiyan Wang, Huangying Zhan, Chunhua Shen, Ming-Ming Cheng, and Ian Reid. Unsupervised scale-consistent depth and ego-motion learning from monocular video. *Advances in neural information processing systems*, 32, 2019.

Patrick Bouthemy, Marc Gelgon, and Fabrice Ganansia. A unified approach to shot change detection and camera motion characterization. *IEEE transactions on circuits and systems for video technology*, 9(7):1030–1044, 1999.

Ling-Yu Duan, Jesse S Jin, Qi Tian, and Chang-Sheng Xu. Nonparametric motion characterization for robust classification of camera motion patterns. *IEEE Transactions on Multimedia*, 8(2):323–340, 2006.

Zhiwen Fan, Jian Zhang, Renjie Li, Junge Zhang, Runjin Chen, Hezhen Hu, Kevin Wang, Peihao Wang, Huaizhi Qu, Shijie Zhou, et al. Vlm-3r: Vision-language models augmented with instruction-aligned 3d reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 31054–31065, 2026.

Haoan Feng, Sri Harsha Musunuri, and Guan-Ming Su. Geometry-guided camera motion understanding in VideoLLMs. *arXiv preprint arXiv:2603.13119*, 2026.

Mickael Guironnet, Denis Pellerin, and Michele Rombaut. Camera motion classification based on transferable belief model. In *2006 14th European signal processing conference*, pp. 1–5. IEEE, 2006.

Yongxin Guo, Jingyu Liu, Mingda Li, Qingbin Liu, Xi Chen, and Xiaoying Tang. Trace: Temporal grounding video llm via causal event modeling. *arXiv preprint arXiv:2410.05643*, 2024.

Muhammad Abul Hasan, Min Xu, Xiangjian He, and Changsheng Xu. CAMHID: Camera motion histogram descriptor and its application to cinematographic shot classification. *IEEE Transactions on Circuits and Systems for Video Technology*, 24(10):1682–1695, 2014.

Qingqiu Huang, Yu Xiong, Anyi Rao, Jiaze Wang, and Dahua Lin. Movienet: A holistic dataset for movie understanding. In *European conference on computer vision*, pp. 709–727. Springer, 2020.

Rebecca Killick, Paul Fearnhead, and Idris A Eckley. Optimal detection of changepoints with a linear computational cost. *Journal of the American Statistical Association*, 107(500):1590–1598, 2012.

Yoon Kim and Alexander M Rush. Sequence-level knowledge distillation. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pp. 1317–1327, 2016.

Zhengqi Li, Richard Tucker, Forrester Cole, Qianqian Wang, Linyi Jin, Vickie Ye, Angjoo Kanazawa, Aleksander Holynski, and Noah Snavely. Megasam: Accurate, fast, and robust structure and motion from casual dynamic videos. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10486–10496. IEEE, 2025.

Zhiqiu Lin, Siyuan Cen, Daniel Jiang, Jay Karhade, Hewei Wang, Chancharik Mitra, Yu Tong Tiffany Ling, Yuhan Huang, Rushikesh Zawar, Xue Bai, et al. Towards understanding camera motions in any video. *Advances in Neural Information Processing Systems*, 38, 2026.

Xinhang Liu, Yu-Wing Tai, and Chi-Keung Tang. ChatCam: Empowering camera control through conversational AI. *Advances in Neural Information Processing Systems*, 37:54483–54506, 2024.

Hugh Christopher Longuet-Higgins and Kvetoslav Prazdny. The interpretation of a moving retinal image. *Proceedings of the royal society of London. Series B, Biological sciences*, pp. 385–397, 1980.

Sang-Cheol Park, Hyoung-Suk Lee, and Seong-Whan Lee. Qualitative estimation of camera motion parameters from the linear composition of optical flow. *Pattern Recognition*, 37(4):767–779, 2004.

Antonia Petrogianni, Panagiotis Koromilas, and Theodoros Giannakopoulos. Film shot type classification based on camera movement styles. In *Iberian Conference on Pattern Recognition and Image Analysis*, pp. 602–615. Springer, 2022.

Chris Rockwell, Joseph Tung, Tsung-Yi Lin, Ming-Yu Liu, David F Fouhey, and Chen-Hsuan Lin. Dynamic camera poses and where to find them. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12444–12455. IEEE, 2025.

Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Dingfeng Shi, Yujie Zhong, Qiong Cao, Lin Ma, Jia Li, and Dacheng Tao. Tridet: Temporal action detection with relative boundary modeling. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18857–18866, 2023.

Yuwen Tan, Joey Huang, Jin Huang, Haoxiang Li, and Boqing Gong. Natural language camera movement understanding. In *European Conference on Computer Vision*, pp. 262–280. Springer, 2026.

Charles Truong, Laurent Oudre, and Nicolas Vayatis. Selective review of offline change point detection methods. *Signal processing*, 167:107299, 2020.

Hee Lin Wang and Loong-Fah Cheong. Taxonomy of directing semantics for film shot classification. *IEEE transactions on circuits and systems for video technology*, 19(10):1529–1542, 2009.

Jiahao Wang, Yufeng Yuan, Rujie Zheng, Youtian Lin, Jian Gao, Lin-Zhuo Chen, Yajie Bao, Chang Zeng, Yanxi Zhou, Xiao-Xiao Long, et al. Spatialvid: A large-scale video dataset with spatial annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 42592–42603, 2026a.

Jianyuan Wang, Minghao Chen, Shangzhan Zhang, Nikita Karaev, Johannes Schönberger, Patrick Labatut, Piotr Bojanowski, David Novotny, Andrea Vedaldi, and Christian Rupprecht. VGGT-Ω. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026b.

Ye Wang, Ziheng Wang, Boshen Xu, Yang Du, Kejun Lin, Zihan Xiao, Zihao Yue, Jianzhong Ju, Liang Zhang, Dingyi Yang, et al. Time-r1: Post-training large vision language model for temporal video grounding. *Advances in Neural Information Processing Systems*, 38:83330–83364, 2026c.

Tianwei Xiong, Yuqing Wang, Daquan Zhou, Zhijie Lin, Jiashi Feng, and Xihui Liu. LVD-2M: A long-take video dataset with temporally dense captions. *Advances in Neural Information Processing Systems*, 37:16623–16644, 2024.

Antoine Yang, Arsha Nagrani, Paul Hongsuck Seo, Antoine Miech, Jordi Pont-Tuset, Ivan Laptev, Josef Sivic, and Cordelia Schmid. Vid2seq: Large-scale pretraining of a visual language model for dense video captioning. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10714–10726. IEEE, 2023.

Songlin Yang, Zhe Wang, Xuyi Yang, Songchun Zhang, Xianghao Kong, Taiyi Wu, Xiaotong Zhao, Ran Zhang, Alan Zhao, and Anyi Rao. ShotVerse: Advancing cinematic camera control for text-driven multi-shot video creation. *arXiv preprint arXiv:2603.11421*, 2026.

Chen-Lin Zhang, Jianxin Wu, and Yin Li. Actionformer: Localizing moments of actions with transformers. In *European conference on computer vision*, pp. 492–510. Springer, 2022.

Mengchen Zhang, Tong Wu, Jing Tan, Ziwei Liu, Gordon Wetzstein, and Dahua Lin. Gendop: Auto-regressive camera trajectory generation as a director of photography. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 18229–18239. IEEE, 2025.

Haoyu Zhao, Zihao Zhang, Jiaxi Gu, Haoran Chen, Qingping Zheng, Pin Tang, Yeyin Jin, Yuang Zhang, Junqi Cheng, Zenghui Lu, et al. CT-1: Vision-language-camera models transfer spatial reasoning knowledge to camera-controllable video generation. *arXiv preprint arXiv:2604.09201*, 2026.

A Data and annotation

A.1 Training data construction and annotation

Source and annotation unit. The training dataset contains approximately 292K distinct video clips from curated film and television footage. Each clip is a temporal crop containing one continuous shot and is paired with a complete camera-motion timeline. Training–benchmark separation is described in Appendix A.2.

Annotation protocol. Annotators follow the shared vocabulary and temporal representation in Sec. 3. Each timeline covers the full clip and records the start and end of each segment together with its set of active motion types, directions, and speeds. Concurrent operations are recorded together, and changes in type, direction, or speed define a new segment. Directions follow the camera's local coordinate system, and speed follows the perceptual tiers defined in the task. The composition illustration in Appendix A.3 and the two-segment example in Appendix B.3 show the representation and its encoding.

Batch organization and reviewer roles. Annotation proceeds in batches, and the composition of subsequent batches is adjusted based on preceding annotation results. The quality-control team comprises four professionally trained reviewers and one additional reviewer responsible for final acceptance. These roles describe the review team; they do not imply that each training clip is independently checked by all five reviewers.

Random checks and revision. Initial quality checks use a 10% random sampling rate and cover motion type, direction, speed, and temporal boundaries. Identified errors are fed back for annotation correction and further training. Batches that fail quality checks are revised and resubmitted for acceptance. When checks reveal elevated or systematic errors, the scope of quality control is expanded and affected data are revised. The sampling rate describes the initial review coverage rather than a measured annotation-accuracy rate.

Relation to benchmark and application review. This initial training-data quality control is separate from the benchmark annotation protocol in Sec. 4.2, under which all three benchmark reviewers independently inspect each shot. It also precedes the geometry-assisted application in Sec. 7.4, which targets selected directed-motion labels while leaving existing speed annotations, hold states, and temporal boundaries unchanged.

A.2 Training–benchmark separation

Training data and TCMU-Bench are separated at the source-video level: no benchmark source video, or frame from such a video, is included in the training data. This applies to the human-annotated training dataset and the student distillation data. Shared-source temporal crops within the benchmark, described below, do not constitute training–benchmark overlap; the benchmark subsets are clip-disjoint rather than source-disjoint.

A.3 Camera-motion vocabulary and composition

Figure 5 illustrates the motion vocabulary, concurrent operations within a segment, and sequential changes across segments.

A.4 Detailed benchmark statistics and curation conditions

Benchmark annotation and quality control. Three annotators with film or photography backgrounds and three reviewers construct the benchmark annotations. Each shot is segmented and labelled by one annotator using the type, direction, and speed vocabulary defined in Sec. 3. All three reviewers independently inspect each shot. Flagged annotations are revised and reviewed again. Remaining label disagreements are adjudicated individually by majority vote among the three annotators and three reviewers. If the vote on any individual label is tied, the entire shot is

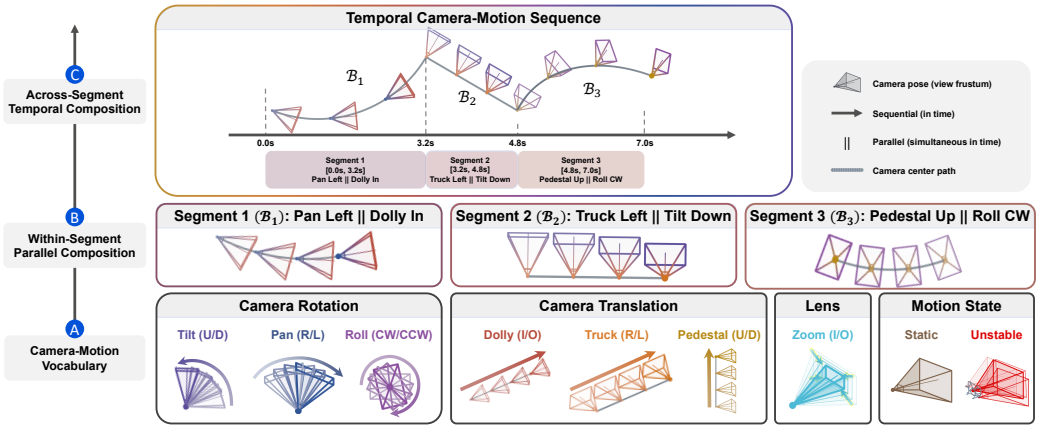


Figure 5: **Hierarchical formulation of temporal camera-motion understanding.** The bottom layer presents the camera-motion vocabulary considered in this work, including camera rotation, camera translation, lens motion, and motion states. The middle layer illustrates within-segment parallel composition, where multiple camera motions may occur simultaneously and are represented as a motion set $\mathcal{B}_i$. The top layer shows across-segment temporal composition, in which consecutive motion sets are arranged along the video timeline to form the complete camera-motion sequence $\mathcal{Y}(V)$. Colored camera frustums denote different motion types, while the camera-center path and temporal boundaries indicate the spatial trajectory and segment transitions, respectively.

excluded. The final 1,000-shot benchmark contains no samples with tied label votes. This benchmark quality-control procedure is separate from the training-annotation review application.

Scale and label distribution. The 2,590 annotated segments contain 4,050 motion-label instances. Of these segments, 903 are Static or Unstable holds, 694 carry one directed motion, and 993 carry multiple directed motions, including 403 with three or more. Shot duration ranges from 1.2 to 19.8s; 900 clips lie between 8 and 20s, and eight are shorter than 2s. Type counts are Pan 650, Tilt 526, Roll 210, Truck 523, Pedestal 418, Dolly 611, Zoom 209, Static 231, and Unstable 672. Zoom in/out contribute 97/112 instances, respectively. Unstable occurs in 541 clips. Figure 2 shows these distributions; most clips have two or three segments, and concurrency ranges from one to five labels. The 23 generated clips total 277.8s and account for 2.3% of TCMU-Bench and 15.0% of the Challenging subset.

We do not balance the numbers of examples across motion types or motion combinations. Motion-type frequencies are shown in Figure 2, and combination coverage is reported in Appendix A.5.

Subset statistics. Segment counts per clip run from one to seven; *easy* reaches three segments and *hard* four, so every longer timeline sits in *challenging* (Fig. 2c). *Easy* holds 1,132 segments and 1,382 labels (1.22 per segment); *hard* 982 / 1,739 (1.77); *challenging* 476 / 929 (1.95). Figure 2d shows one or two labels per segment in *easy*; segments with three or more labels occur only in *hard* and *challenging*. Mean duration is comparable across subsets (12.0–12.7s). The proportion of GT label-frames assigned to `Static` or `Unstable` is 21.5% in *easy*, 10.1% in *hard*, and 7.8% in *challenging*, measured over the full annotated range at 0.1s intervals. These differences contribute to differences in aggregate recognition performance (Appendix D.3).

A.5 Combination coverage of the released benchmark

A *combination* here is the set of motion types co-occurring inside one segment, taken direction-agnostically and with `Static`/`Unstable` excluded, so that `Dolly in+Pan left` and `Dolly out+Pan right` are the same cell. Segments partition across the cells; clips do not, since a clip enters every combination one of its segments realizes. Table 6 lists the single- and two-motion cells,

Table 5: The *challenging* subset by curation condition, grouped into the three families reused in Sec. 7.2. Each clip carries exactly one condition. Confusion and same-direction compounds are pooled across motion pairs, matching the released labels. *No stable background* pools clips whose background cannot lock the camera—empty texture, or subject motion that reads as the camera's. #Labels counts motion instances over the closed set of Sec. 3; a `Static` or `Unstable` stretch contributes one.

Condition	Clips	Segments	Labels
Parallax discriminations			
Confusable-pair confusion	29	67	93
Same-direction compound	11	29	54
`Dolly zoom`	8	27	63
subtotal	48	123	210
Reconstruction-challenging conditions			
No stable background	16	36	45
Camera too close to the subject	11	30	53
Picture-in-picture	11	29	52
Visual effects	7	22	27
subtotal	45	117	177
Structural complexity			
Complex compound movement	39	125	332
More than four segments	21	111	210
subtotal	60	236	542
Challenging, total	153	476	929

which are the ones a closed vocabulary of seven motion types leaves few enough of to fill exhaustively, and Table 7 summarizes the higher cardinalities.

Thirty-eight *easy* clips are temporal crops of videos whose full takes also appear in *hard* or *challenging*. These follow from the *easy* criterion of Sec. 4: when a take runs one configuration and then another, the stretch holding a single configuration forms an *easy* clip and the full take stays in *hard*, where the switch is what the subset is meant to test. Clip identifiers are unique, so the subsets of Sec. 4 remain clip-disjoint; they are not source-disjoint.

A.6 The geometry-assisted annotation review

TCMU-Geo is run over the complete training dataset to check for missing or conflicting motion labels within the existing human segments. Candidates are ranked by the strength and consistency of their geometric evidence, and the most pronounced subset—spanning approximately 2.3% of the training videos—is selected for review. Every candidate is verified by a human before any annotation change is applied; Geo never relabels the training dataset automatically. Human reviewers assign speeds to newly added motions. Existing speed annotations, hold states, and temporal boundaries remain unchanged.

B Methods and training

B.1 Geometric readout illustration and channel mapping

A reconstructed camera admits two different readouts. Its pose increments describe physical motion in translation and rotation units, whereas re-projecting the reconstructed scene describes the displacement written onto the image plane. TCMU-Geo uses the latter because camera-motion annotations are based on visible image motion.

B.2 TCMU-VLM training and released student

8B input and optimization settings. Videos are sampled at 5 fps with 4–100 frames, at most approximately $100{,}352$ pixels per frame, and a maximum sequence length of $16{,}384$ tokens. The

Table 6: Single- and two-motion combination coverage. All 7 single-motion and all 21 two-motion cells are populated. Clips are counted once per cell they enter, so the column sums past 1,000; segments do not repeat.

Combination	Clips	Segments	Easy	Hard	Chal.
Dolly	115	124	50	44	21
Tilt	111	121	78	20	13
Pan	109	124	46	40	23
Truck	97	109	53	29	15
Zoom	83	90	53	20	10
Pedestal	70	73	39	25	6
Roll	46	53	18	18	10
Pan+Truck	59	67	19	19	21
Pedestal+Tilt	55	58	15	32	8
Pan+Tilt	47	54	15	21	11
Dolly+Pan	44	47	16	17	11
Dolly+Truck	41	43	10	22	9
Dolly+Pedestal	40	42	15	19	6
Dolly+Tilt	35	35	18	11	6
Dolly+Roll	31	34	17	10	4
Pan+Zoom	30	32	19	8	3
Pedestal+Truck	30	34	14	13	3
Pan+Pedestal	26	26	13	11	2
Tilt+Zoom	25	25	19	5	1
Tilt+Truck	17	17	12	4	1
Pan+Roll	16	17	7	6	3
Roll+Truck	15	16	7	6	2
Truck+Zoom	13	13	8	2	3
Roll+Tilt	9	9	6	3	0
Pedestal+Roll	6	6	4	2	0
Pedestal+Zoom	6	7	2	1	3
Dolly+Zoom	5	5	0	0	5
Roll+Zoom	3	3	1	0	2

Table 7: Coverage by number of concurrent motions. "Possible" is $\binom{7}{k}$ over the seven directional types of Sec. 3; the $k = 0$ row holds the Static/Unstable stretches, which are not combinations. Segments sum to the 2,590 of Sec. 4.

Concurrent motions k	Cells present	Possible	Segments
0 (filler only)	–	–	903
1	7	7	694
2	21	21	590
3	27	35	344
4	15	35	54
5	1	21	5

prompt states the closed vocabulary and output rules compactly, and requests only JSON. Both 8B models are trained for two epochs with a global batch size of approximately 256, a learning rate of 10^{-5}, bfloat16, and ZeRO data parallelism. The vision encoder, vision-language aligner, and language model are all updated. Comparisons with TCMU-Geo use the original-annotation 8B model; the reviewed 8B model is used for CameraBench evaluation and student distillation. Both versions are reported in Table 1 and compared in Sec. 7.4.

Zero-shot VLM input settings. Gemini-3-flash, Qwen3-VL-8B, and InternVL3.5-8B target 5 fps with a 100-frame cap; LLaVA-Video-7B uses the same target frame rate with a 64-frame limit. For LLaVA-Video-7B, only the first 64 sampled frames are retained when the limit is exceeded, truncating later video content. Evaluation still covers the full shot. All VLMs use the shared system prompt.

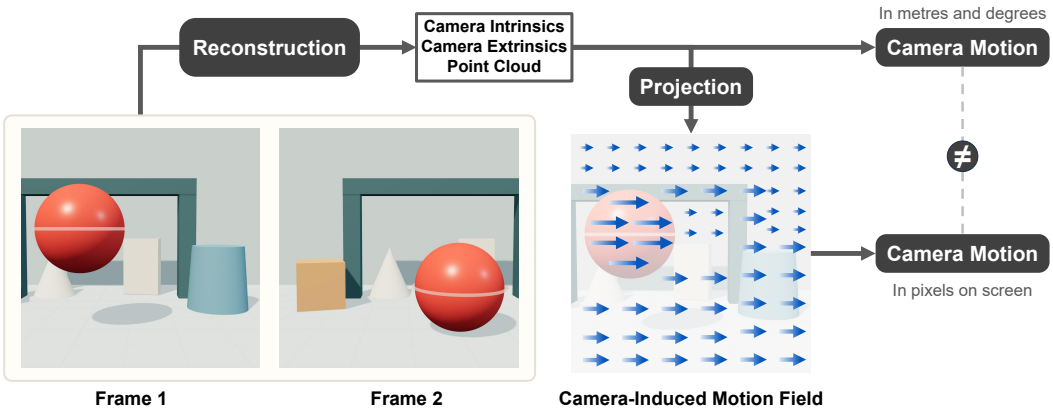


Figure 6: **Camera pose and induced image motion are different readouts of the same reconstruction.** Directly reading the extrinsics describes camera motion in physical translation and rotation units. Re-projecting the reconstructed scene instead produces the image-plane displacement induced by that motion. In this leftward truck example, nearby points on the ball move much farther on screen than points on the distant background. This depth dependence is absent from the pose magnitude but provides the parallax signal used by TCMU-Geo to distinguish translation from rotation. The camera-induced motion field is synthesized by reprojecting reconstructed scene points, rather than estimated from observed pixel motion.

Table 8: Camera-induced motion-field component and basis coefficient mapped onto the closed set. The same divergence coefficient is `Zoom` in $\mathbf{f}_{\text{rot}}$ and `Dolly` in $\mathbf{f}_{\text{trans}}$.

Stream	Coefficient	Type
$\mathbf{f}_{\text{rot}}$	a	`Pan`
$\mathbf{f}_{\text{rot}}$	b	`Tilt`
$\mathbf{f}_{\text{rot}}$	d	`Roll`
$\mathbf{f}_{\text{rot}}$	c	`Zoom`
$\mathbf{f}_{\text{trans}}$	a	`Truck`
$\mathbf{f}_{\text{trans}}$	b	`Pedestal`
$\mathbf{f}_{\text{trans}}$	c	`Dolly`
$\mathbf{f}$	median magnitude	`Static`
$\mathbf{f}_{\text{trans}}$	fit residual	`Unstable`

B.2.1 Released TCMU-VLM-2B

We distill the reviewed 8B teacher into a Qwen3-VL-2B student using teacher-generated structured sequences (Kim & Rush, 2016). Source videos from LVD-2M (Xiong et al., 2024) are split into continuous shots and labelled by the teacher using the same prompt and frame-sampling settings. After removing overlap with TCMU-Bench, we select 200K clips to balance motion types as far as the teacher labels permit. The student trains for two epochs on these teacher-generated sequences, without direct access to the human annotations, using the teacher's learning rate, batch size, maximum sequence length, and unfrozen vision components. Its benchmark results are reported in Table 1.

B.3 JSON output schema

Annotations and model predictions use the JSON schema below as an interchange format for $\mathcal{Y}(V)$; the task itself is the segment sequence of Sec. 3, not this encoding. Listing 1 shows a two-segment example whose second segment is compound.

Listing 1: Example TCMU output. The second segment shows a compound motion (`Dolly in + Pan left`).

```
{
  "segments": [
    {
      "start_time": 0.0,
      "end_time": 2.5,
      "camera_motions": [
        {"type": "Pan", "direction": "left", "speed": "medium"}
      ]
    },
    {
      "start_time": 2.5,
      "end_time": 5.0,
      "camera_motions": [
        {"type": "Dolly", "direction": "in", "speed": "slow"},
        {"type": "Pan", "direction": "left", "speed": "slow"}
      ]
    }
  ]
}
```

B.4 TCMU-Geo configuration

TCMU-Geo uses the 1B VGGT-Ω model at input resolution 512 and samples frames at approximately 10 fps. Rates are measured in frame-width-normalized displacement per second. All motion channels exceeding their respective thresholds are retained.

Table 9: Key hyperparameters used for TCMU-Geo evaluation.

Setting	Value
Minimum segment duration	0.5s
Change-point penalty	2.5
Static magnitude threshold	0.010
Pan / Tilt / Roll thresholds	0.020 / 0.020 / 0.030
Truck / Pedestal / Dolly thresholds	0.020 / 0.010 / 0.020
Zoom threshold	0.060
Unstable residual / minimum magnitude	0.6 / 0.020
Channel selection	All channels above threshold
Translation slow maximum / fast minimum	0.03 / 0.20
Rotation slow maximum / fast minimum	0.10 / 0.30
Zoom slow maximum / fast minimum	0.10 / 0.30
Merge adjacent identical labels	enabled

The Zoom channel is derived from divergence in the depth-independent stream, with direction assigned from focal-length change.

C Evaluation and reproducibility

C.1 Detailed evaluation protocol

All systems are evaluated against the same ground truth (GT) under the following protocol. Equivalent label encodings are mapped to common identifiers. Empty predictions and uncovered intervals are empty label sets. Unknown labels and other predicted labels unsupported by GT count as false positives, including unsupported co-occurrences of hold states and directed motions.

Evaluation domain and preprocessing. Each shot is evaluated over its complete annotated range $[0, T)$. Before scoring, shot-endpoint alignment sets the first segment's start time to 0 and the last segment's end time to T for each nonempty predicted timeline, leaving all internal boundaries unchanged. During scoring, intervals are clipped to this range without further extension. Valid timelines contain finite, positive-duration, non-overlapping intervals. Adjacent identical states are merged when extracting events and boundaries.

Frame recognition. Predicted and annotated label timelines are sampled at 0.1s intervals, independently of the model's input frame rate. At $t = 0, 0.1, \ldots < T$, let G_t and P_t be sets of type–direction labels. We accumulate

$$TP = \sum_t |G_t \cap P_t|, \qquad FP = \sum_t |P_t \setminus G_t|, \qquad FN = \sum_t |G_t \setminus P_t|.$$

Sums also run across videos. Precision, recall and F1 are $TP/(TP + FP)$, $TP/(TP + FN)$ and $2TP/(2TP + FP + FN)$. Micro scores are computed after pooling counts; a zero denominator yields an undefined score, displayed as a dash. The sampling grid covers the full GT range, and intervals are left-closed and right-open. No duration filter is applied, so a sub-0.1s motion contributes only if it contains a sampled timestamp. The full benchmark contains 122,497 sample points and 200,146 positive label-frame instances. The type view collapses every direction, including Dolly and Zoom in/out. GT concurrency uses the full type–direction set.

Speed. Grouping by annotated speed measures type–direction recall at each GT speed, independently of predicted speed. Strict-speed evaluation instead uses complete type–direction–speed tuples on the same sampling grid. A speed mismatch contributes one FP tuple and one FN tuple.

Motion-event localization. For each label independently, consecutive intervals on which that label remains active are merged on both GT and prediction sides. Changes in another concurrent label or in speed do not split a type–direction event, and gaps in the label's activity separate events. All positive-duration events participate. Within each video and label, a GT event g and predicted event p are eligible for matching when

$$\mathrm{tIoU}(g, p) = \frac{|g \cap p|}{|g \cup p|} \geq \theta, \qquad \theta \in \{0.3, 0.5, 0.7\}.$$

One-to-one matching first maximizes the number of matches and then their total tIoU. Matched pairs, unmatched predictions, and unmatched annotations count as TP, FP, and FN. The type–direction view contains 3,382 GT events; the type view used in paired analysis contains 3,249. The same event sets are evaluated at all thresholds.

State-boundary localization. We extract internal timestamps where the set of complete type–direction–speed tuples changes. Transitions into or out of an empty predicted region count as changes; shot endpoints 0 and T are excluded. Predicted and annotated boundaries are matched one-to-one within absolute error $\delta \in \{0.25, 0.5, 1.0\}$ seconds. Matching first maximizes the number of matches and then minimizes their summed absolute error. Matched, extra, and missed boundaries count as TP, FP, and FN. Matching compares timestamps without requiring agreement of adjacent motion labels. The same 1,590 GT boundaries are evaluated at each tolerance. A video with no predicted or annotated boundary contributes no TP/FP/FN counts.

Paired recovery and confusion. Paired analysis fixes the GT type-view events and compares whether each model's independent matching recovers each event. The four outcomes are both, neither, only Geo, and only VLM. For frame confusion, a matched type contributes one unit to the diagonal. On the recall side, each missing GT type distributes one unit uniformly across active predicted types, or to `none` if there is no prediction. On the precision side, each unsupported predicted type distributes one unit uniformly across active GT types, or to `none` if the GT set is empty. The recall-side matrix is conditioned on GT types and the precision-side matrix on predicted types; each is row-normalized over all columns, including `none`.

D Complete experimental results

D.1 Full event, boundary and speed results

The following tables report P/R/F1 and support for each evaluated system. Event counts and boundary counts differ because events are extracted separately for each motion label, whereas a complete-state transition contributes one global boundary.

Table 10: Motion-event localization at tIoU 0.3 (%; type–direction). All positive-duration events participate.

System	P	R	F1	GT units	Predicted units
SpatialVID	32.0	32.1	32.1	3,382	3,392
TCMU-Geo	51.5	59.1	55.0	3,382	3,883
Gemini-3-flash (zero-shot)	27.0	22.3	24.4	3,382	2,790
Qwen3-VL-8B (zero-shot)	36.3	13.0	19.1	3,382	1,205
InternVL3.5-8B (zero-shot)	11.6	5.4	7.4	3,382	1,584
LLaVA-Video-7B (zero-shot)	5.8	3.9	4.7	3,382	2,251
TCMU-VLM-8B (original)	81.7	64.1	71.9	3,382	2,654
TCMU-VLM-8B (reviewed)	83.8	67.8	74.9	3,382	2,735
TCMU-VLM-2B	76.1	59.9	67.0	3,382	2,664
TCMU-VLM-8B (vision frozen)	74.7	64.5	69.2	3,382	2,923
TCMU-Geo (MegaSaM)	37.8	56.8	45.4	3,382	5,084

Table 11: Motion-event localization at tIoU 0.5 (%; type–direction). All positive-duration events participate.

System	P	R	F1	GT units	Predicted units
SpatialVID	21.8	21.9	21.8	3,382	3,392
TCMU-Geo	45.8	52.6	48.9	3,382	3,883
Gemini-3-flash (zero-shot)	18.7	15.4	16.9	3,382	2,790
Qwen3-VL-8B (zero-shot)	29.2	10.4	15.3	3,382	1,205
InternVL3.5-8B (zero-shot)	7.3	3.4	4.7	3,382	1,584
LLaVA-Video-7B (zero-shot)	2.8	1.9	2.2	3,382	2,251
TCMU-VLM-8B (original)	79.6	62.5	70.0	3,382	2,654
TCMU-VLM-8B (reviewed)	80.8	65.3	72.2	3,382	2,735
TCMU-VLM-2B	73.2	57.7	64.5	3,382	2,664
TCMU-VLM-8B (vision frozen)	71.2	61.5	66.0	3,382	2,923
TCMU-Geo (MegaSaM)	32.6	49.0	39.1	3,382	5,084

D.2 Comparisons with shared reconstruction and motion vocabulary

We compare SpatialVID with two TCMU-Geo variants using MegaSaM and VGGT-Ω. The MegaSaM variant reuses the same reconstruction outputs as SpatialVID; motion-field decomposition, thresholds, temporal decoding, and endpoint processing are unchanged from the default VGGT-Ω configuration. All variants are evaluated on the same $1{,}000$ shots under the protocol in Sec. 6.

Over the full nine-type vocabulary, TCMU-Geo with MegaSaM achieves 53.9 frame F1, 39.1 Event-F1@0.5, and 33.3 Boundary-F1@0.5s, compared with 32.8, 21.8, and 18.3 for SpatialVID. The default VGGT-Ω variant achieves higher frame and event F1 (60.7 and 48.9), while its boundary F1 is slightly lower (32.5).

We also evaluate the seven shared types: Pan, Tilt, Roll, Truck, Pedestal, Dolly, and Static, retaining direction distinctions for thirteen labels. Zoom and Unstable are removed from both GT and predictions while all shots and complete GT time domains are retained. Frame evaluation uses the same 0.1s grid, including timestamps whose GT labels become empty. Events are extracted after filtering without bridging temporal gaps. Under this shared vocabulary, TCMU-Geo with MegaSaM reaches 58.5 frame F1 and 43.5 Event-F1@0.5 versus 36.4 and 25.0 for SpatialVID. Its advantage holds at all three event thresholds (Table 17), showing that the gains persist with both reconstruction outputs and vocabulary matched.

D.3 Recognition without hold-state labels

Table 18 compares Geo and the original VLM after excluding Static and Unstable from GT and predictions. On Easy, Geo's frame F1 changes from 57.0 over the full vocabulary to 63.8 over directed labels; the VLM changes from 79.0 to 77.5. This comparison shows how hold-state errors contribute to the aggregate scores.

Table 12: Motion-event localization at tIoU 0.7 (%; type–direction). All positive-duration events participate.

System	P	R	F1	GT units	Predicted units
SpatialVID	10.2	10.3	10.2	3,382	3,392
TCMU-Geo	34.8	39.9	37.2	3,382	3,883
Gemini-3-flash (zero-shot)	10.8	8.9	9.8	3,382	2,790
Qwen3-VL-8B (zero-shot)	18.7	6.7	9.8	3,382	1,205
InternVL3.5-8B (zero-shot)	4.0	1.9	2.6	3,382	1,584
LLaVA-Video-7B (zero-shot)	1.5	1.0	1.2	3,382	2,251
TCMU-VLM-8B (original)	75.2	59.0	66.1	3,382	2,654
TCMU-VLM-8B (reviewed)	75.5	61.0	67.5	3,382	2,735
TCMU-VLM-2B	67.6	53.3	59.6	3,382	2,664
TCMU-VLM-8B (vision frozen)	63.4	54.8	58.7	3,382	2,923
TCMU-Geo (MegaSaM)	24.6	37.0	29.5	3,382	5,084

Table 13: State-boundary localization at 0.25s tolerance (%). GT support is fixed across tolerances.

System	P	R	F1	GT units	Predicted units
SpatialVID	8.2	13.9	10.3	1,590	2,707
TCMU-Geo	18.5	16.6	17.5	1,590	1,424
Gemini-3-flash (zero-shot)	9.0	7.4	8.1	1,590	1,313
Qwen3-VL-8B (zero-shot)	8.8	0.5	1.0	1,590	91
InternVL3.5-8B (zero-shot)	8.4	3.1	4.5	1,590	582
LLaVA-Video-7B (zero-shot)	4.4	6.9	5.3	1,590	2,491
TCMU-VLM-8B (original)	88.2	75.1	81.1	1,590	1,353
TCMU-VLM-8B (reviewed)	77.3	67.0	71.8	1,590	1,379
TCMU-VLM-2B	73.7	62.8	67.8	1,590	1,355
TCMU-VLM-8B (vision frozen)	58.6	54.9	56.7	1,590	1,491
TCMU-Geo (MegaSaM)	16.3	22.3	18.9	1,590	2,173

D.4 Complete results across benchmark subsets

Tables 19 and 20 expand Table 2 with additional event thresholds and boundary tolerances. Every comparison uses complete shots and all concurrency levels under the protocol in Sec. 6.

D.5 Paired event-recovery details

The GT type-view event set is identical for the two methods and remains fixed at every tIoU threshold. Table 21 uses $\theta = 0.5$. Geo's directed-event recall is 66.0% versus 56.5% for the original VLM; for holds the ordering reverses, 11.5% versus 80.2%.

Table 14: State-boundary localization at 0.5s tolerance (%). GT support is fixed across tolerances.

System	P	R	F1	GT units	Predicted units
SpatialVID	14.5	24.7	18.3	1,590	2,707
TCMU-Geo	34.4	30.8	32.5	1,590	1,424
Gemini-3-flash (zero-shot)	18.4	15.2	16.7	1,590	1,313
Qwen3-VL-8B (zero-shot)	30.8	1.8	3.3	1,590	91
InternVL3.5-8B (zero-shot)	17.0	6.2	9.1	1,590	582
LLaVA-Video-7B (zero-shot)	9.6	15.0	11.7	1,590	2,491
TCMU-VLM-8B (original)	92.4	78.6	84.9	1,590	1,353
TCMU-VLM-8B (reviewed)	87.1	75.5	80.9	1,590	1,379
TCMU-VLM-2B	84.7	72.2	78.0	1,590	1,355
TCMU-VLM-8B (vision frozen)	73.9	69.3	71.5	1,590	1,491
TCMU-Geo (MegaSaM)	28.9	39.4	33.3	1,590	2,173

Table 15: State-boundary localization at 1.0s tolerance (%). GT support is fixed across tolerances.

System	P	R	F1	GT units	Predicted units
SpatialVID	23.1	39.4	29.1	1,590	2,707
TCMU-Geo	52.8	47.3	49.9	1,590	1,424
Gemini-3-flash (zero-shot)	31.6	26.1	28.6	1,590	1,313
Qwen3-VL-8B (zero-shot)	42.9	2.5	4.6	1,590	91
InternVL3.5-8B (zero-shot)	32.0	11.7	17.1	1,590	582
LLaVA-Video-7B (zero-shot)	14.1	22.1	17.3	1,590	2,491
TCMU-VLM-8B (original)	93.9	79.9	86.4	1,590	1,353
TCMU-VLM-8B (reviewed)	91.3	79.2	84.8	1,590	1,379
TCMU-VLM-2B	88.7	75.6	81.6	1,590	1,355
TCMU-VLM-8B (vision frozen)	81.6	76.5	79.0	1,590	1,491
TCMU-Geo (MegaSaM)	43.6	59.6	50.4	1,590	2,173

Table 16: Strict type–direction–speed frame P/R/F1 (%; 0.1s sampling).

System	P	R	F1
SpatialVID	3.3	2.6	2.9
TCMU-Geo	37.1	43.2	39.9
Gemini-3-flash (zero-shot)	18.1	15.4	16.6
Qwen3-VL-8B (zero-shot)	17.7	12.0	14.3
InternVL3.5-8B (zero-shot)	7.6	4.7	5.8
LLaVA-Video-7B (zero-shot)	6.8	3.8	4.9
TCMU-VLM-8B (original)	69.8	54.8	61.4
TCMU-VLM-8B (reviewed)	70.2	56.8	62.8
TCMU-VLM-2B	65.0	49.8	56.4
TCMU-VLM-8B (vision frozen)	59.8	50.3	54.6
TCMU-Geo (MegaSaM)	30.2	37.8	33.6

Table 17: Reconstruction-backend and vocabulary comparisons (%). All configurations use the same 1,000 shots. Frame P/R/F1 use type–direction labels at 0.1s sampling, Event-F1 uses the listed tIoU thresholds, and Boundary-F1 uses 0.5s tolerance. SpatialVID and TCMU-Geo (MegaSaM) reuse identical reconstruction outputs. Shared-vocabulary scores exclude Zoom and Unstable from GT and predictions while retaining the full time domain; boundary scores are reported only for the full vocabulary.

System	Frame P	R	F1	Event@0.3	@0.5	@0.7	Boundary@0.5s
Full nine-type vocabulary							
SpatialVID (MegaSaM)	37.5	29.2	32.8	32.1	21.8	10.2	18.3
TCMU-Geo (MegaSaM)	48.5	60.7	53.9	45.4	39.1	29.5	33.3
TCMU-Geo (VGGT-Ω)	56.4	65.8	60.7	55.0	48.9	37.2	32.5
Shared seven-type vocabulary							
SpatialVID (MegaSaM)	37.5	35.4	36.4	36.7	25.0	11.7	–
TCMU-Geo (MegaSaM)	48.5	73.6	58.5	50.4	43.5	32.9	–
TCMU-Geo (VGGT-Ω)	56.2	76.7	64.9	61.2	54.7	41.9	–

Table 18: Frame F1 over directed labels, excluding Static and Unstable (%; 0.1s sampling). Other unsupported predicted labels remain false positives.

System	All	Easy	Hard	Challenging
TCMU-Geo	65.7	63.8	67.9	65.4
TCMU-VLM-8B (original)	69.6	77.5	67.7	57.6

Table 19: Complete-shot recognition and temporal localization for TCMU-Geo (%). Frame uses type–direction labels at 0.1s sampling.

Stratum	Frame F1	Event@0.3	@0.5	@0.7	Boundary@0.25s	@0.5s	@1.0s
All	60.7	55.0	48.9	37.2	17.5	32.5	49.9
Easy	57.0	48.6	46.1	38.4	15.7	29.0	46.6
Hard	63.8	59.5	50.6	36.2	18.0	33.7	51.5
Challenging	62.6	58.8	51.3	36.8	20.1	36.8	53.0
Parallax discriminations	60.8	58.0	52.2	35.9	22.5	36.6	49.3
Reconstruction-challenging	55.8	46.3	41.7	28.6	13.9	25.0	37.5
Structural complexity	67.4	64.6	55.1	40.8	21.7	42.2	61.5

Table 20: Complete-shot recognition and temporal localization for original-annotation TCMU-VLM (%). Frame uses type–direction labels at 0.1s sampling.

Stratum	Frame F1	Event@0.3	@0.5	@0.7	Boundary@0.25s	@0.5s	@1.0s
All	71.5	71.9	70.0	66.1	81.1	84.9	86.4
Easy	79.0	80.1	79.7	78.3	89.7	93.2	94.1
Hard	68.8	68.3	65.2	60.2	78.5	81.7	83.5
Challenging	59.5	61.2	58.7	51.9	68.1	74.0	75.8
Parallax discriminations	51.8	61.3	59.2	51.4	67.2	77.9	80.9
Reconstruction-challenging	63.7	65.0	64.2	56.9	56.0	60.8	65.6
Structural complexity	62.2	59.6	56.4	50.2	73.8	77.9	77.9

Table 21: Per-type paired GT-event recovery at tIoU 0.5. Directions are collapsed before event extraction. Counts are descriptive and do not assume independent events.

Group	GT events	Geo R	VLM R	Both	Neither	Only Geo	Only VLM
Pan	481	70.7	54.3	210	90	130	51
Tilt	435	64.4	48.3	170	115	110	40
Roll	159	69.2	63.5	74	22	36	27
Truck	363	72.7	54.3	169	71	95	28
Pedestal	307	70.0	58.3	146	59	69	33
Dolly	417	66.9	53.2	169	85	110	53
Zoom	184	33.2	84.2	51	19	10	104
Static	231	42.0	61.9	66	57	31	77
Unstable	672	1.0	86.5	7	91	0	574

D.6 TYPE-LEVEL CONFUSION ANALYSIS



Figure 7: Recall-side type confusion at 0.1s sampling. Rows are GT types; columns are predictions including none. Each row sums to 100%; n is its GT label-frame support. Missing-label mass is shared across active predicted types.

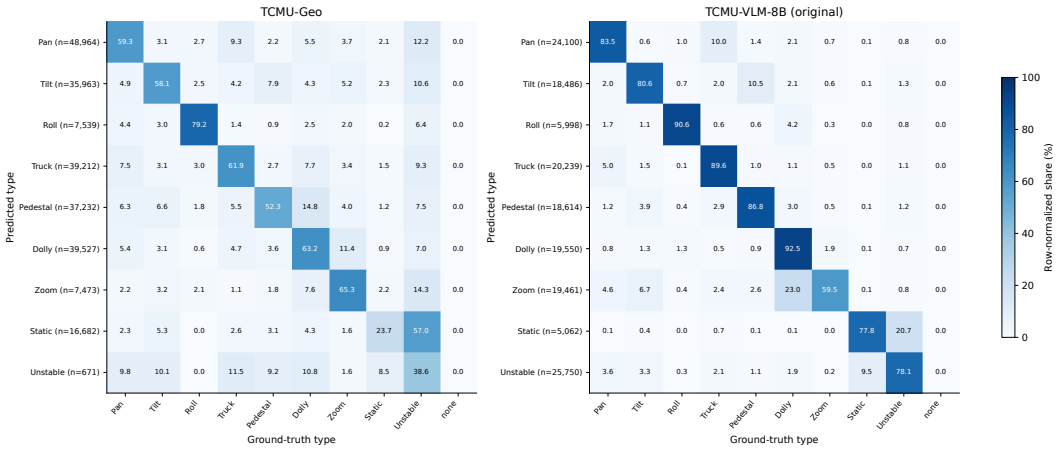


Figure 8: Precision-side type confusion at 0.1s sampling. Rows are predicted types; columns are GT types including none. Each row sums to 100%; n is predicted label-frame support. Unsupported-label mass is shared across active GT types.

Geo has higher type recall for the rotational and translational motions, whereas the original VLM has much higher recall for Zoom and Unstable. Dolly/Zoom errors are asymmetric: the VLM's Dolly-to-Zoom share is 20.6%, while Geo's Zoom-to-Dolly share is 24.1%. Geo assigns 43.1% of GT Unstable mass to Static; the VLM assigns 37.8% of GT Static mass to Unstable. Each percentage is normalized by its source row, not by the number of errors alone.

D.7 CAMERABENCH PROTOCOL AND CATEGORY-LEVEL RESULTS

Camera Motion Recognition on CameraBench. CameraBench (Lin et al., 2026) provides an external 15-category recognition benchmark over 969 clips. We collapse each predicted timeline into a clip-level label set and map our vocabulary to its categories; `Unstable`, which has no counterpart, is ignored. Because our structured output does not provide the continuous scores used by CameraBench's average-precision protocol, we compare F1. TCMU-VLM uses a parameter-

free label-presence decision, whereas CameraBench-SFT uses a per-category threshold selected to maximize F1 on the test set.

Table 22: CameraBench binary classification over 15 categories and 969 clips. TCMU-VLM uses the reviewed checkpoint; CameraBench-SFT uses per-category thresholds selected on the test set.

			TCMU-VLM			CameraBench-SFT		
Category	n	pos	P	R	F1	P	R	F1
Move_In	549	161	89.8	82.0	**85.7**	62.3	82.0	70.8
Move_Out	549	53	76.9	75.5	**76.2**	42.9	56.6	48.8
Move_Up	430	29	90.9	34.5	**50.0**	27.0	34.5	30.3
Move_Down	430	37	85.7	48.6	**62.1**	52.4	59.5	55.7
Move_Right	487	77	93.3	72.7	**81.8**	44.4	71.4	54.7
Move_Left	488	56	91.8	80.4	**85.7**	36.4	64.3	46.5
Zoom_In	450	50	93.9	62.0	**74.7**	41.3	62.0	49.6
Zoom_Out	452	46	100.0	69.6	**82.1**	59.3	69.6	64.0
Pan_Right	512	77	89.9	92.2	**91.0**	70.1	61.0	65.3
Pan_Left	512	79	98.6	89.9	**94.0**	63.5	77.2	69.7
Tilt_Up	458	58	91.5	74.1	**81.9**	65.4	58.6	61.8
Tilt_Down	456	35	90.6	82.9	**86.6**	70.0	40.0	50.9
Roll_Clockwise	470	42	100.0	38.1	**55.2**	31.2	69.0	43.0
Roll_Counterclockwise	470	48	96.0	50.0	**65.8**	37.5	75.0	50.0
Static	966	94	64.7	95.7	**77.3**	54.7	50.0	52.2
Macro			90.3	69.9	**76.7**	50.5	62.0	54.2

D.8 Annotation-review results by subset

Table 23 reports frame precision, recall, and F1 before and after annotation review for each benchmark subset and Challenging condition family.

Table 23: Frame P/R/F1 before and after annotation review on the same benchmark (%; type–direction, 0.1s sampling). Deltas are computed before rounding.

Stratum	Original P	R	F1	Reviewed P	R	F1	ΔF1
All	81.2	63.8	71.5	84.3	68.2	75.4	4.0
Easy	83.4	75.0	79.0	86.1	78.2	82.0	2.9
Hard	82.4	59.0	68.8	84.4	63.9	72.8	4.0
Challenging	72.8	50.3	59.5	79.6	56.2	65.9	6.4
Parallax discriminations	61.9	44.5	51.8	77.7	57.8	66.3	14.5
Reconstruction-challenging	68.2	59.9	63.7	75.1	63.2	68.6	4.9
Structural complexity	84.0	49.3	62.2	83.7	52.0	64.1	2.0

D.9 Recognition by concurrency and annotated speed

Table 24 measures type–direction recall within annotated speed tiers, not speed prediction. The fast tier contains only eight segment-label annotations. Table 25 reports type–direction precision, recall, and F1 for Geo and the original VLM by GT concurrency. The original VLM has higher recall for

Table 24: Recall by annotated speed tier (%; type–direction, 0.1s sampling). Support counts GT label-frames.

GT speed	GT label-frames	Geo R	VLM R	ΔR
zero	6,472	61.1	60.9	0.2
slow	93,460	51.0	57.8	-6.8
medium	99,978	79.8	69.6	10.2
fast	236	78.0	59.3	18.6

Table 25: Frame P/R/F1 by GT concurrency (%; type–direction, 0.1s sampling). The VLM uses the original annotations.

GT concurrency	GT label-frames	Geo P	R	F1	VLM P	R	F1
1	68,218	35.6	53.3	42.7	79.9	82.0	81.0
2	68,134	67.1	72.4	69.7	78.4	60.1	68.1
3	52,059	78.4	72.0	75.1	87.4	49.5	63.2
4	10,240	84.8	72.0	77.8	91.8	44.1	59.5
5	1,495	99.3	71.8	83.4	83.1	35.5	49.8

slow motions, whereas Geo has higher recall for medium motions and increasing concurrency. The fast-tier result remains limited by its small sample.

D.10 Vision-side fine-tuning ablation

Both runs train for two epochs on the original annotations, using the same training configuration except for whether the vision components are frozen. Unfreezing improves frame F1 by 1.7 points, Event-F1@0.5 by 4.0, and Boundary-F1@0.5s by 13.4. Frame recall changes from 64.2% to 63.8%, so the frame improvement is driven by precision rather than recall.

Table 26: Vision-side fine-tuning ablation under all three metrics (%). Frame uses 0.1s sampling; Event uses tIoU thresholds and Boundary uses 0.5s tolerance.

System	Frame P	R	F1	Event@0.3	@0.5	@0.7	Boundary@0.5s
Vision frozen	76.4	64.2	69.8	69.2	66.0	58.7	71.5
Vision unfrozen	81.2	63.8	71.5	71.9	70.0	66.1	84.9

E Qualitative examples

E.1 When the VLM is stronger

Figures 9 and 10 illustrate selected complementary outcomes from the predicted timelines. Both compare Geo with the original-annotation VLM. A method can have better recognition but worse boundary timing on the same clip.



Figure 9: **Examples favoring the annotation-trained VLM.** Rows show GT, TCMU-Geo, and the original-annotation VLM. Black text denotes correctly recognized labels, orange text extra predictions, and gray strikethrough text missed GT labels added for comparison. The markings summarize qualitative type–direction differences rather than exact frame-level or event-matching decisions; speed does not determine text color. L/R, U/D, I/O, and CW/CCW denote left/right, up/down, in/out, and clockwise/counterclockwise; S/M/F/Z denote slow/medium/fast/zero. The examples illustrate omitted, substituted, and unsupported motion components.

Figure 9 highlights two recurring limitations of the geometric pipeline. Under reconstruction difficulty, residual errors can activate unsupported motion channels or obscure hold states. Longer sequences introduce a second challenge: the geometric decoder may merge or fragment successive motion configurations. The VLM better preserves the dominant semantic states and their ordering by using appearance and temporal context. These cases illustrate its recognition advantage under reconstruction-challenging conditions and its stronger temporal localization under structural complexity (Table 2).

E.2 When the geometric solver is stronger



Figure 10: Examples favoring TCMU-Geo on geometrically confusable and compound motion. The camera-induced motion-field decomposition separates operations with similar appearance but different depth dependence and preserves concurrent components that the VLM may omit or substitute. Rows show ground truth, TCMU-Geo, and the original-annotation VLM. Orange text denotes unsupported predicted labels, while gray strikethrough text denotes missed GT labels added for comparison, not model outputs. These markings summarize type–direction differences; speed does not determine text color. Abbreviations follow Figure 9.

Figure 10 illustrates complementary geometric recoveries. The VLM omits a component or substitutes an operation with similar appearance. Geo has higher frame F1 on these selected clips, but not uniformly higher boundary F1: in the first clip the VLM has more accurate complete-state boundary timing, despite missing motion components. These examples illustrate why component recognition and state-boundary localization are evaluated separately; they do not establish a general geometric advantage within every condition family.

E.3 ANNOTATION-REVIEW EXAMPLES

Figure 11 shows two clips in which the reviewed VLM recovers omitted components or removes unsupported labels. These examples illustrate recognition improvements; the aggregate boundary result prevents interpreting them as uniform temporal improvement.

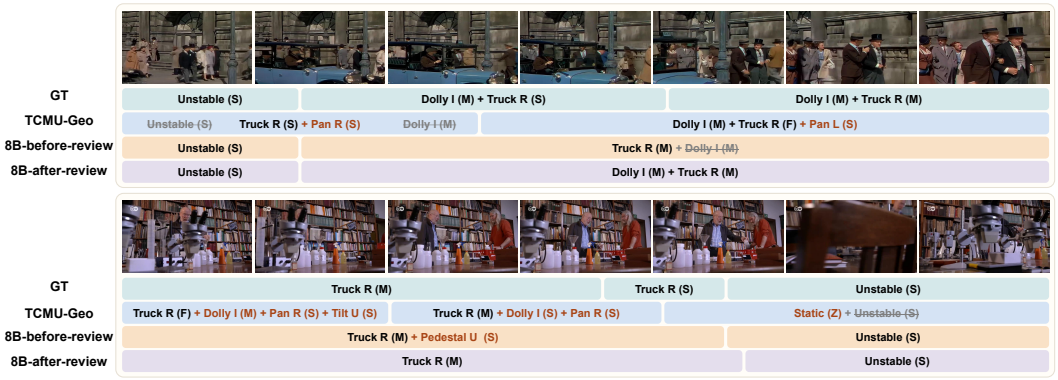


Figure 11: **Annotation-review examples.** The reviewed VLM recovers an omitted motion component in the first clip and removes an unsupported component in the second. In prediction rows, black text denotes correctly recognized motion labels, orange text denotes extra predictions, and gray strikethrough text denotes missed GT labels added for comparison, not model outputs. These annotations summarize type–direction differences qualitatively; they do not encode exact frame-level or event-matching decisions. Parentheses show predicted speed for model outputs and annotated speed for missed labels; speed does not determine text color. Abbreviations follow Figure 9.

F LIMITATIONS AND FUTURE WORK

TCMU is defined on continuous shots and requires external shot-boundary detection for edited videos. Recognition and temporal localization remain imperfect, particularly for concurrent motions. TCMU-Geo depends on estimated cameras and depth and can fail when reconstruction is unreliable; its thresholds also imperfectly distinguish holds, shake, and weak directed motion. TCMU-VLM depends on the coverage and quality of its training annotations. Rare combinations and fast motions have small support, limiting conclusions about arbitrary motion compositions.

Frame and event metrics score type–direction labels, while the state-boundary metric scores change timing without checking the labels across matched boundaries. Supplementary strict-speed recognition assesses speed assignment. Short events may be missed by the 0.1s frame grid while remaining eligible for event and boundary scoring. Reported performance is measured after shot-endpoint alignment (Appendix C.1); raw shot-endpoint prediction accuracy remains unevaluated. Future work should improve speed and boundary estimation and broaden coverage of rare motions.

Geometry-assisted human review improves frame and event F1 but reduces boundary F1, despite unchanged boundary annotations. This application evaluates the effect of the reviewed annotations on the trained model; it does not establish the relative efficiency of different candidate-selection strategies.