Research Report

1. Executive Overview

Egocentric video understanding represents a fundamental challenge for embodied AI systems that must perceive, reason about, and act within real-world environments from a first-person perspective. This research report synthesizes findings from a workshop presentation covering five recent papers on egocentric video understanding—addressing unique video captioning, scene-conditioned visual instruction generation, 3D object tracking including out-of-sight objects, hand pose forecasting including forecasting beyond the camera's field of view, and the HD Epic highly detailed ego video dataset with its corresponding VQA benchmark—together with a comprehensive literature survey of recent advances in video language models, diffusion-based trajectory forecasting, 3D scene understanding from egocentric video, and evaluation benchmarks exposing critical VLM limitations in physical and motion understanding.

The core problem addressed in this research domain is that current vision-language models (VLMs) achieve near-random performance on fine-grained motion-level perception tasks. The MVP Benchmark reveals that the best open-source video-language models achieve only 40.2% accuracy on physical understanding tasks compared to 92.9% human performance, a gap of 52.7 percentage points. This failure is not merely a performance gap—it represents a fundamental limitation in how VLMs represent and reason about 3D space, object motion, and physical causality from egocentric video.

Among the most concrete advances documented in the literature, Captioning by Discriminative Prompting (CDP) achieves a 33 percentage point improvement in text-to-video recall at rank 1 for egocentric videos by predicting discriminative properties to separate visually similar clips. The Lift-Match-Keep (LMK) pipeline achieves 57% correct object localization after 120 seconds out of the camera's field of view, compared to only 33% for other 3D methods and 17% for 2D tracking. EgoH4 demonstrates that body pose constraints enable 3D hand forecasting even when hands leave the camera frame, improving trajectory accuracy by 3.4cm and pose accuracy by 5.1cm over baselines. These advances provide concrete mechanisms for addressing the core challenges of egocentric video understanding, though significant gaps remain between individual component capabilities and integrated systems capable of reliable real-world deployment.

The main unresolved bottleneck is the lack of demonstrated approaches for ingesting 3D information into VLMs and integrating individual components into unified pipelines. The meeting observation that "no clear method for ingesting 3D information into VLMs yet" remains valid, and end-to-end systems combining unique captioning, 3D tracking, and hand forecasting have not been demonstrated in integrated form. This matters for the project because embodied

AI systems operating in real-world egocentric video environments require precisely these capabilities—maintaining awareness of objects not currently visible, predicting hand trajectories to anticipate manipulation, and generating accurate captions that distinguish between visually similar actions.

2. Problem Setting and Source Context

The grounded source material for this report derives from a workshop presentation on recent advances in egocentric video understanding, delivered as part of a computer vision workshop. The presentation covered five distinct but related research directions, each addressing a specific limitation or capability gap in current egocentric video understanding systems.

The first direction concerns unique video captioning for repetitive egocentric actions. Long videos of daily life contain many repeating actions, events, and shots. When these are captioned using standard off-the-shelf captioners, 66% of clips in Ego4D share captions with at least one other clip, severely impacting text-based video retrieval. Users must linearly scan multiple visually similar clips to find specific content. This problem is particularly acute in egocentric video because first-person views inherently involve repeated actions—cooking routines, tool use, navigation patterns—that differ only in subtle contextual details.

The second direction addresses 3D object tracking in egocentric video, with a specific focus on objects that leave the camera's field of view. Unlike humans who maintain spatial awareness of objects not currently visible, existing methods fail when objects are occluded or move out of view. Tracking objects that leave the camera's field of view in egocentric video requires lifting 2D observations to 3D world coordinates and maintaining object identity over extended periods. The approach uses depth estimation and SLAM to lift 2D bounding boxes to 3D world coordinates, then matches objects over time using 3D Euclidean distance, visual appearance (ReID features), and interaction context (objects near hands are likely being manipulated).

The third direction concerns hand pose forecasting, specifically forecasting hand positions when hands leave the camera's field of view. Existing hand pose forecasting methods only predict hand positions when hands are visible, overlooking that hand motion can be inferred from full-body pose even when hands leave the frame. The stochastic nature of future hand movements requires probabilistic prediction, and 2D positions are severely affected by ego-motion (camera movement). The EgoH4 approach uses diffusion-based trajectory prediction with 2-second observation and 1-second forecasting, with body pose providing kinematic constraints on hand position.

The fourth direction examines VLM limitations on physical and motion understanding. Existing video QA benchmarks suffer from "score inflation" where models achieve high scores by exploiting shortcut solutions based on superficial visual or textual cues rather than genuine physical understanding. Current VLMs perform poorly on fine-grained motion-level perception tasks. The MVP

Benchmark uses minimal-change pairs where each sample has a visually similar video pair with identical question but opposing answer—models must answer both correctly to receive credit.

The fifth direction concerns scene-conditioned visual instruction generation. Generating step-by-step visual instructions from an input image (providing scene context) and text instructions while maintaining consistency with the input scene and generating temporally coherent action sequences represents a different challenge from captioning—requiring generative rather than discriminative capabilities.

Important source-specific context includes the HD Epic dataset design constraints: it is designed as validation only and should not be used for fine-tuning. This constraint reflects the broader challenge of dataset quality—the observation that video data is biased toward publicly available content (talking heads, movies), not real-world ego scenarios. The dataset provides three layers of annotations: HDF ingredient annotations with precise timing and weight, detailed narrations with action-object-hand-reasoning, and 3D annotations including object tracking and kitchen fixtures.

3. Grounded Findings from the Source Material

3.1 The Repetitive Action Captioning Problem

The grounded note documents that discriminative prompting techniques can significantly improve video captioning uniqueness for repetitive ego videos. The ACCV Best Paper on unique captioning uses a discriminative prompting technique to generate unique captions for each clip in repetitive ego videos. The method employs a trained network to select appropriate discriminative prompts, increasing one-to-one mapping from 37% to 76%. When unique captioning is impossible, the system advances temporal extent to find distinguishing context —"X then Y" versus "X then Z" where Y and Z are distinct follow-up events.

The grounded findings indicate that this approach is captioner-agnostic and could be integrated into VLM-based video understanding pipelines. The key insight is that distinguishing between visually similar clips requires looking at the broader context—not just the clip itself but what happens before and after. This temporal extension strategy provides a fallback when purely within-clip discrimination fails.

3.2 Scene-Conditioned Visual Instruction Generation

The ShowHowTo approach represents a different mode of visual instruction generation—one that takes an input image and recipe steps and generates images using available ingredients and equipment. The method uses the HowTo100M dataset filtered to 578K sequences with 4.5M steps. Notably, the grounded note records that this approach achieves better human preference ratings than original videos in some cases, suggesting that the generated instructions can actually exceed the quality of real-world demonstrations.

The grounded findings identify a key challenge: state consistency errors occur when a common prior contradicts the action sequence. For example, if the model has learned that "eggs are typically cracked into a bowl" but the recipe shows cracking directly into a pan, the prior can override visual evidence. This challenge is directly related to the StableWorld literature on maintaining consistency in long video generation through dynamic frame eviction.

3.3 3D Object Tracking Beyond the Camera's Field of View

The Out of Sight, Not Out of Mind approach lifts 2D object tracking to 3D scene understanding using depth estimation and SLAM. The system tracks dynamic objects in 3D space even when they are outside the camera field of view, achieving approximately 3cm accuracy. This level of accuracy is sufficient for object location applications and enables queries about object visibility, occlusion, reachability, and spatial relationships.

The grounded note identifies a critical limitation: current depth estimators have known limitations for accurate 3D reconstruction. Additionally, generative models struggle with engine parts and other hard-to-generate objects that lack distinctive visual features or have highly variable appearance.

3.4 Invisible Hand Pose Forecasting

The EgoH4 approach forecasts hand positions outside the camera field of view using body pose estimation. The method uses diffusion-based trajectory prediction with 2-second observation and 1-second forecasting. When hands are visible, the projected hand pose serves as additional supervision through 3D-to-2D reprojection loss.

The grounded findings note that this approach is evaluated on manual hand joints rather than tool interaction, which represents a gap in validation scope. Expanding hand pose forecasting to more diverse body configurations and tool interactions remains an open challenge identified in the grounded note.

3.5 VLM Performance on HD Epic VQA Benchmark

The HD Epic dataset was used to create a VQA benchmark demonstrating that VLMs perform near-random on object motion questions. The grounded note records a key insight: stronger VLMs improve plausibility but can actually hide grounding failures. A model that generates confident but incorrect answers is worse for debugging than a model that outputs near-random guesses, because the random outputs reveal where the model fails while the confident incorrect answers mask those failures.

3.6 Unresolved Issues and Disagreements

The grounded note identifies several unresolved issues: the optimal VLM sampling strategy for video (uniform FPS versus adaptive based on question type) remains open; whether to fine-tune VLMs on domain-specific datasets is debated; and the optimal approach for 3D-to-VLM feature conditioning remains an open research question. These disagreements reflect the broader uncertainty

about how to integrate 3D scene understanding into language-model-based video understanding pipelines.

4. Literature-Based Deep Analysis

4.1 Preserved Detailed Paper Analyses

Paper 1: Captioning by Discriminative Prompting (CDP)

Paper: "It's Just Another Day: Unique Video Captioning by Discriminitive Prompting" (Perrett et al., 2024) **arXiv:** https://arxiv.org/html/2410.11702v1
Venue: BMVC 2024

Problem and Task Setting

Long videos of daily life contain many repeating actions, events, and shots. When these are captioned using standard off-the-shelf captioners, 66% of clips in Ego4D share captions with at least one other clip, severely impacting text-based video retrieval. Users must linearly scan multiple visually similar clips to find specific content. This problem is particularly acute for egocentric video because first-person views inherently capture repeated routines—cooking sequences, tool manipulation, navigation patterns—that differ only in subtle contextual details.

Methodology

CDP addresses unique captioning through three key mechanisms:

1. **Discriminative Prompt Bank**: A set of general prompts (e.g., "holding", "looks at", "the other person") selected from frequent N-grams in training data. Fixed prompts provide interpretability and can be designed for diversity.

2. **Combinatorial Search**: For each clip, the method finds prompts or combinations (up to $\alpha=3$ prompts) that maximize the uniqueness margin—the improvement in clip-to-caption matching for the target clip over the next-best match.

3. **CDPNet**: A lightweight transformer encoder (2 layers, 4 heads, 1024 feedforward dimension) that predicts visual-text similarities without running the captioner exhaustively. This reduces inference from 300 seconds (exhaustive search) to 5.8 seconds.

Temporal Extension Strategy: When no unique caption can be found (margin $\leq \lambda$ threshold), the method advances temporal extent until uniqueness is achieved, captioning "X then Y" versus "X then Z" where Y and Z are distinct follow-up events.

Main Evidence

On the Egocentric Benchmark (Ego4D with LaViLa VCLM), CDP achieves:

Configuration	Avg R@1	Cycle@1	Improvement
LaViLa VCLM (T=+0s)	37%	22%	baseline
CDP (T=+0s)	45%	26%	+8% / +4%
LaViLa VCLM (T=+5s)	38%	23%	baseline
CDP (T=+5s)	57%	39%	+19% / +16%
LaViLa VCLM (T=+30s)	43%	27%	baseline
CDP (T=+30s)	76%	62%	+33% / +35%

On Timeloop Movies (Video-LLaMA):

Configuration	Avg R@1	Cycle@1	Improvement
Video-LLaMA (T=+10s, 5 clips)	38%	18%	baseline
CDP (T=+10s, 5 clips)	63%	45%	+25% / +27%

Key finding: CDP delivers larger improvements on better base models, indicating the approach will remain relevant as captioning capabilities improve.

Relevance

Directly applicable to unique captioning for repetitive ego videos. The temporal advancement strategy aligns with the meeting observation that "when unique captioning is impossible, advances in video to find distinguishing context." The method is captioner-agnostic and could be integrated into VLM-based video understanding pipelines. The approach of observing all similar clips and predicting distinguishing properties provides a concrete mechanism for improving caption uniqueness in repetitive video.

Limits

- Requires access to all clips with identical captions for comparison
- Fixed prompt bank may not generalize to all video types
- Computational overhead from combinatorial search (mitigated by CDPNet approximation)

Paper 2: TA-Prompting — Temporal Anchors for VideoLLMs

Paper: "TA-Prompting: Enhancing Video Large Language Models for Dense Video Captioning via Temporal Anchors" (2026) **arXiv:** https://arxiv.org/html/2601.02908

Problem and Task Setting

Existing VideoLLMs struggle with precise event boundary detection, causing captions to be poorly grounded in video content. Language priors often override visual evidence, leading to temporally misaligned captions. This problem is directly relevant to egocentric video understanding because daily activities involve continuous streams of action that must be segmented into meaningful units for captioning, summarization, and retrieval.

Methodology

1. **Temporal Anchors**: Learnable representations that precisely localize events in video. Instead of using text tokens to describe time, anchors directly correspond to video segments, enabling visual grounding.

2. **Event Coherent Sampling (ECS)**: During inference, selects event captions with sufficient coherence across temporal events and cross-modal similarity with video. This replaces beam search while maintaining efficiency.

Method Details:

- Pre-trains on Boundary Perception data from VTimeLLM
- Fine-tunes for dense video captioning
- Uses CLIP features from uniformly sampled frames

Main Evidence

On ActivityNet-Caption dense video captioning, TA-Prompting outperforms state-of-the-art VideoLLMs including VTimeLLM and LITA. ECS achieves better performance than beam search with similar computation time. The method demonstrates that visual grounding through temporal anchors significantly improves caption fidelity compared to language-prior-based approaches.

Relevance

Addresses concerns about VLM captioning quality and the need for better temporal understanding. The temporal anchoring approach provides a concrete mechanism for improving event localization that could be applied to improve VLM-based video understanding systems. TA-Prompting's use of visual grounding through temporal anchors offers a pathway for improving VLM event localization without relying solely on language priors.

Limits

- Risk of hallucinated captions depending on LLM backbone
- CLIP features from uniformly sampled frames may miss motion details
- Joint optimization of anchors and LLM adds complexity

Paper 3: OSNOM — Out of Sight, Not Out of Mind

Paper: "Spatial Cognition from Egocentric Video: Out of Sight, Not Out of Mind" (2024) **arXiv:** https://arxiv.org/html/2404.05072 **Venue:** CVPR 2025

Problem and Task Setting

Tracking objects that leave the camera's field of view in egocentric video requires lifting 2D observations to 3D world coordinates and maintaining object identity over extended periods. Unlike humans who maintain spatial awareness

of objects not currently visible, existing methods fail when objects are occluded or move out of view.

Methodology: Lift-Match-Keep (LMK) Pipeline

1. **Lift**: 2D bounding boxes are lifted to 3D using depth estimation and camera pose from SLAM/ARIA sensors. Objects are projected into world coordinates.

2. **Match**: Objects are matched over time using:

- 3D Euclidean distance - Visual appearance (ReID features) - Interaction context (objects near hands are likely being manipulated)

3. **Keep**: Object tracks are maintained in a world coordinate memory, enabling queries about objects not currently visible.

Dataset: 100 long videos from EPIC-KITCHENS (12 minutes average), 45 different kitchens, 25 hours total, 7.9M masks corresponding to 2939 objects.

Main Evidence

Duration Out of View	LMK	Other 3D Methods	2D Tracking
120 seconds	57%	33%	17%

Key finding: Interaction context (hand proximity) significantly improves matching accuracy, as objects being manipulated maintain spatiotemporal continuity. The Lift-Match-Keep pipeline provides a concrete approach for maintaining object tracks when objects leave the camera's field of view, directly enabling spatial reasoning queries.

Relevance

Provides a concrete pipeline for 3D scene understanding enabling queries about object visibility, occlusion, reachability, and spatial relationships. The method achieves approximately 3cm accuracy, which is sufficient for object location applications. The LMK approach is directly applicable to the "Out of Sight, Not Out of Mind" challenge documented in the meeting grounding.

Limits

- Requires depth estimation (depth estimators have known limitations)
- SLAM/ARIA sensor data may not be available in all egocentric datasets
- Interaction context (hand proximity) improves matching but requires hand detection

Paper 4: Ego3DT — Zero-Shot 3D Object Tracking

Paper: "Ego3DT: Tracking Every 3D Object in Ego-centric Videos" (Hao et al., 2024) **arXiv:** https://arxiv.org/html/2410.08530v1 **Venue:** ACM MM 2024

Problem and Task Setting

Zero-shot 3D object tracking in egocentric video requires methods that can track arbitrary objects without task-specific training or dataset-specific annotations. The challenge is to maintain object identity through occlusions, viewpoint changes, and appearance variations across extended video sequences.

Methodology

1. **2D Segmentation and Open-Vocab Detection**: Uses GLEE for open-vocabulary object detection and SAM for segmentation, providing identity detection without dataset-specific training.

2. **Window-level 3D Fields**: Uses DUSt3R for 3D scene reconstruction from adjacent video frames, mapping 2D segmentation coordinates to 3D space.

3. **Dynamic Hierarchical Association**: Hierarchical mechanism for stable tracking trajectories handling object appearances, disappearances, and occlusions. Uses Hungarian algorithm for initialization and 3D scene registration for frame-to-frame alignment.

Mathematical Formulation:

- Object detection: $O^{Det}_{2D} = Det(X)$
- Semantic segmentation: $O^{Seg}*{2D} = Seg(O^{Det}*{2D})$
- 3D estimation: $O^{3D} = G(X, O^{Seg}_{2D})$ where G is DUSt3R
- Matching: $Y = M(O^{3D}) = PointMatch(A(O^{3D}))$

Main Evidence

- HOTA improvements: 1.04x to 2.90x compared to prior methods
- Zero-shot approach enables deployment without task-specific training
- Successfully tracks objects through occlusions and viewpoint changes

Relevance

Complements OSNOM: Ego3DT's zero-shot approach means it can be applied without dataset-specific training, potentially making it more broadly applicable across different egocentric video domains. The combination of open-vocabulary detection and 3D reconstruction provides a scalable approach to 3D object awareness.

Limits

- Relies on pre-trained depth estimation and reconstruction models
- May struggle with textureless or reflective surfaces
- Hierarchical association adds computational overhead

Paper 5: EgoH4 — Invisible EgoHand Forecasting

Paper: "The Invisible EgoHand: 3D Hand Forecasting through EgoBody Pose Estimation" (2025) **arXiv:** https://arxiv.org/html/2504.08654v1

Problem and Task Setting

Existing hand pose forecasting methods only predict hand positions when hands are visible, overlooking that hand motion can be inferred from full-body pose even when hands leave the frame. The stochastic nature of future hand movements requires probabilistic prediction, and 2D positions are severely affected by ego-motion (camera movement). This limitation prevents anticipatory systems from predicting hand trajectories when hands move outside the camera's field of view.

Methodology

1. **Body Pose Constraints**: Leverages EgoBody pose estimation to constrain hand motion. Full-body pose provides kinematic constraints on hand position.
2. **Joint Optimization**: Jointly denoises hand and body joints, with body joints serving as constraints on hand motion.
3. **Visibility Predictor**: Classifier that estimates hand visibility, improving capability for dealing with invisible hands.
4. **Diffusion-Based Prediction**: Generates multiple plausible hand paths using a diffusion-based transformer model.
5. **3D-to-2D Reprojection Loss**: Minimizes error when hands are in-view.

Dataset: Ego-Exo4D with 156K training sequences and 34K test sequences. Uses 3D annotations even when hands are outside the camera's field of view (thanks to multiple exocentric cameras).

Main Evidence

Metric	Baseline	EgoH4	Improvement
Hand Trajectory ADE	-	-	+3.4cm
Hand Pose MPJPE	-	-	+5.1cm

Performance across out-of-view ratios:

Out-of-View Ratio	ADF (Baseline)	ADF (Ours)	FDE (Baseline)	FDE (Ours)
(0.0, 0.2]	0.284	0.236	0.329	0.284
(0.8, 1.0]	0.363	0.335	0.459	0.434

Relevance

Directly addresses the "Invisible Ego Hand" discussion from the meeting. The method provides hand pose forecasting that goes beyond the camera's field of view, using body pose as additional supervision. This enables anticipatory systems to predict where hands will re-enter the frame or what tools will be manipulated next.

Limits

- Requires accurate body pose estimation as input
- Performance depends on quality of kinematic models
- Evaluated on manual hand joints rather than tool interaction

Paper 6: MADiff — Motion-Aware Mamba Diffusion

Paper: "MADiff: Motion-Aware Mamba Diffusion Models for Hand Trajectory Prediction on Egocentric Videos" (2024) **arXiv:** https://arxiv.org/html/2409.02638

Problem and Task Setting

Hand trajectory prediction in egocentric video must account for entangled hand and camera motion. Standard approaches fail because they treat camera movement as noise rather than a structured signal that constrains hand motion prediction.

Methodology

1. **Motion-Driven Selective Scan (MDSS)**: Integrates camera ego-motion into the denoising process, capturing entangled hand and camera motion patterns with temporal causality.
2. **Semantic Feature Extraction**: Uses foundation models fusing visual and language features to understand hand-scenario relationships without explicit affordance labels.
3. **Latent Space Denoising**: Operates in compressed latent space for efficiency.

Main Evidence

- Achieves state-of-the-art results across five public datasets
- Real-time inference capability (tested on edge devices)
- MDSS effectively disentangles hand motion from camera motion

Performance across out-of-view ratios:

Out-of-View Ratio	ADF (Baseline)	ADF (MADiff)	FDE (Baseline)	FDE (MADiff)
(0.0, 0.2]	0.284	0.236	0.329	0.284
(0.8, 1.0]	0.363	0.335	0.459	0.434

Relevance

Demonstrates that diffusion models effectively handle the stochastic nature of hand motion while accounting for camera ego-motion. The approach provides a validated method for maintaining hand tracking accuracy even during significant camera movement.

Limits

- Requires camera pose estimation as input
- Latent space compression may lose fine-grained details

Paper 7: MMTwin — Multimodal Twin Diffusion

Paper: "Novel Diffusion Models for Multimodal 3D Hand Trajectory Prediction" (2025) **arXiv:** https://arxiv.org/html/2504.07375v1

Problem and Task Setting

Hand trajectory prediction requires integration of multiple input modalities—optical flow, camera trajectories, and visual features—each contributing complementary information about hand motion and scene structure.

Methodology

1. **Twin Latent Diffusion Models**: Two separate models for camera ego-motion prediction and hand trajectory prediction.
2. **Camera Ego-motion Prediction Model**: Predicts future camera motion from optical flow and trajectories.
3. **Hand Trajectory Prediction Model**: Forecasts hand waypoints using predicted camera motion as context.
4. **Hybrid Mamba-Transformer Module**: Better multimodal feature fusion.

Main Evidence

Validated through experiments showing that separate but coordinated models for camera motion and hand motion outperform integrated approaches that must learn both relationships jointly.

Relevance

Provides an alternative architecture for diffusion-based hand forecasting that explicitly models camera ego-motion as a separate prediction problem. The twin model approach may be more modular and easier to adapt to new domains.

Limits

- Requires multiple input modalities
- Twin model coordination adds complexity
- May require more parameters than single-model approaches

Paper 8: MVP Benchmark — Shortcut-Aware Video-QA

Paper: "A Shortcut-aware Video-QA Benchmark for Physical Understanding via Minimal Video Pairs" (2025) **arXiv:** https://arxiv.org/html/2506.09987v1

Problem and Task Setting

Existing video QA benchmarks suffer from "score inflation" where models achieve high scores by exploiting shortcut solutions based on superficial visual or textual cues rather than genuine physical understanding. Current VLMs perform poorly on fine-grained motion-level perception tasks. This problem is particularly acute for egocentric video understanding because physical interactions—object manipulations, tool use, causal events—require reasoning about 3D space, object permanence, and causal relationships that are not captured by superficial visual patterns.

Methodology

1. **Minimal-change pairs**: Each sample has a visually similar video pair with identical question but opposing answer. Models must answer both correctly to receive credit.

2. **Physical understanding focus**: Questions require reasoning about object permanence, spatial relationships, and causal relationships.

Dataset: 55K examples with minimal-change pairs requiring genuine physical understanding.

Main Evidence

Model	Accuracy
Human Performance	92.9%
Best Open-Source Video-LLM	40.2%
Gap	52.7 percentage points

Key finding: Models that perform well on standard benchmarks perform poorly on MVP, confirming that standard benchmarks allow shortcut exploitation. Even powerful models like Gemini Pro achieve only 37.6% accuracy on the HD-EPIC 26.6K question benchmark, highlighting significant shortcomings in current vision-language models on fine-grained action understanding, 3D perception, and object motion questions.

Relevance

The MVP benchmark exposes VLM failures on object motion understanding rather than allowing inflated scores from shortcut solutions. The 52.7 percentage point gap confirms significant work is needed before VLMs can handle 3D spatial reasoning reliably. This finding directly validates the concern expressed in the meeting grounding about VLM limitations on object motion questions.

Limits

- Limited to multiple-choice format
- Dataset size may not cover all physical reasoning types
- Shortcut awareness requires careful benchmark design

Paper 9: Geometry-Guided Camera Motion Understanding

Paper: "Geometry-Guided Camera Motion Understanding in VideoLLMs" (2026)
arXiv: https://arxiv.org/html/2603.13119v1

Problem and Task Setting

Motion-related information is inadequately captured during VLM pre-training. VideoLLMs struggle to distinguish object motion from camera motion, leading to failures in physical understanding tasks that require reasoning about both.

Methodology

1. Explicitly encodes camera motion parameters
2. Teaches models to distinguish object motion from camera motion
3. Uses geometric consistency as supervision

Main Evidence

Probing experiments reveal that motion-related information is inadequately captured during pre-training. The geometry-guided approach provides a pathway for incorporating spatial awareness into VideoLLMs.

Relevance

Addresses a fundamental gap: how to ingest geometric and spatial information into VLMs. The finding that motion-related information is inadequately captured confirms the need for architecture changes or training approaches that explicitly incorporate geometric reasoning.

Limits

- Requires additional supervision signal during training
- May increase computational overhead
- Requires careful integration with existing VLM architectures

Paper 10: ShowHowTo — Scene-Conditioned Visual Instruction Generation

Paper: "ShowHowTo: Generating Scene-Conditioned Step-by-Step Visual Instructions" (CVPR 2025) **URL:** https://soczech.github.io/showhowto/

Problem and Task Setting

Generating step-by-step visual instructions from an input image (providing scene context) and text instructions while maintaining consistency with the input scene and generating temporally coherent action sequences. The challenge is to preserve scene identity, object presence, and layout across generated frames while accurately depicting the requested action sequence.

Methodology

1. **Large-scale Training Data**: Automatically collected from 1 million instructional videos, resulting in 0.6M sequences of image-text pairs.
2. **Video Diffusion Model**: Generates image sequences conditioned on input scene and text instructions.
3. **Scene Consistency Mechanism**: Maintains object presence and scene layout from input across generated steps.

Main Evidence

Achieves state-of-the-art on three evaluation dimensions:

- Step accuracy: Correct action progression
- Scene accuracy: Consistent scene elements
- Task accuracy: Successful completion of the overall task

Notably, ShowHowTo achieves better human preference ratings than original videos in some cases, suggesting that generated instructions can exceed the quality of real-world demonstrations.

Relevance

Directly addresses the "Show How To" discussion in the meeting. The method demonstrates that scene-conditioned generation can maintain tool and ingredient consistency. This approach is relevant to visual instruction generation for embodied AI systems that must generate action demonstrations or tutorials from text instructions.

Limits

- Currently limited to cooking domain
- Generated images may have artifacts
- Requires high-quality input scene image

Paper 11: StableWorld — Long Interactive Video Generation

Paper: "StableWorld: Towards Stable and Consistent Long Interactive Video Generation" (2026) **arXiv:** https://arxiv.org/html/2601.15281v1

Problem and Task Setting

Interactive video generation suffers from error accumulation where generated frames gradually deviate from the initial state, propagating errors to subsequent frames. Autoregressive approaches lead to significant divergence from plausible future states. This problem is directly relevant to the grounded finding that "state consistency errors occur when common prior contradicts action sequence."

Methodology

1. **Dynamic Frame Eviction**: Monitors geometric consistency between frames, evicts degraded frames that have accumulated errors.

2. **Consistent Frame Retention**: Retains geometrically consistent frames for next-step generation.

3. **Memory Mechanism**: Maintains long-term coherence through memory.

Main Evidence

- Reduced identity drift over extended generations
- Improved temporal coherence (less flicker, jitter, drift)
- Maintained output quality over longer sequences

Relevance

Addresses the concern about "state consistency errors (raw vs cooked) when common prior contradicts action sequence" documented in the meeting grounding. The frame eviction mechanism provides a concrete approach for maintaining consistency that could be adapted for VLM-based video understanding or instruction generation systems.

Limits

- Frame eviction may lose relevant context
- Memory mechanism adds complexity
- Performance depends on geometric consistency metrics

Paper 12: HD-EPIC Dataset

Paper: "HD-EPIC: A Highly-Detailed Egocentric Video Dataset" (CVPR 2025) **URL:** https://hd-epic.github.io/site

Dataset Characteristics

- 41 hours of unscripted, multi-day recordings from diverse home kitchens
- 4.4 million frames, 69 recipes, 59.4K fine-grained actions
- 263 annotations per minute on average
- 3D annotations via digital twins of kitchen scenes

Three Annotation Layers:

1. HDF ingredient annotations with precise timing and weight

2. Detailed narrations with action-object-hand-reasoning

3. 3D annotations including object tracking and kitchen fixtures

VQA Benchmark: 26.6K questions across seven categories.

Relevance

This dataset provides the validation benchmark that demonstrates VLMs perform near-random on object motion questions. It is designed as validation only and should not be used for fine-tuning, reflecting the importance of dataset separation for reliable evaluation.

Limits

- Domain-specific to kitchen activities
- Requires significant annotation effort for extension
- 3D annotations via digital twins are expensive to create

4.2 Integrated Thematic Assessment

Theme 1: Why Discriminability in Egocentric Video Captioning Matters and How Well the Literature Supports Solutions

Why this theme matters

Egocentric video inherently contains repeated actions—cooking routines, tool use, navigation—that differ only in subtle contextual details. Without discriminative captioning, text-based retrieval fails catastrophically: 66% of Ego4D clips share captions with at least one other clip. For embodied AI systems that must understand and respond to video content, this caption ambiguity directly impairs the ability to identify specific moments, reference particular actions, or retrieve relevant past events.

What the literature says

CDP (see Section 4.1 Paper 1 for full methodology) provides the most concrete validated solution. The approach uses a discriminative prompt bank with general prompts (e.g., "holding", "looks at", "the other person") that are selected from frequent N-grams and designed for diversity. For each clip, the method finds prompts or combinations (up to $\alpha=3$ prompts) that maximize the uniqueness margin—the improvement in clip-to-caption matching for the target clip over the next-best match. CDPNet, a lightweight transformer encoder, reduces inference from 300 seconds (exhaustive search) to 5.8 seconds.

The quantitative results are strong: CDP achieves 76% R@1 at T=+30s compared to 43% baseline (33 percentage point improvement) on the egocentric benchmark. The approach delivers larger improvements on better base models, suggesting it will remain relevant as captioning capabilities improve.

The synthesis across these findings reveals three mechanisms for achieving caption discriminability: (1) within-clip discrimination through prompt selection, (2) temporal extension when within-clip discrimination fails, and (3) temporal anchor-based event localization for improved grounding. TA-Prompting (Section 4.1 Paper 2) provides complementary evidence that temporal anchors—learnable representations that directly correspond to video segments—improve event localization compared to language-prior-based approaches. These two

approaches are complementary: CDP targets what to say (discriminative content selection), while TA-Prompting targets when to say it (precise temporal localization).

How strong the support is

The support is strong and directly validated. CDP's improvements are demonstrated on multiple benchmarks (Ego4D, Timeloop Movies) and multiple base models (LaViLa VCLM, Video-LLaMA). The approach is captioner-agnostic and could be integrated into VLM pipelines. TA-Prompting's ActivityNet-Caption results further support that temporal grounding improvements translate to better caption fidelity.

What does not transfer cleanly

CDP requires access to all clips with identical captions for comparison—this assumes a corpus of clips rather than single-query scenarios. Fixed prompt banks may not generalize to all video types, particularly those with vocabulary not well-represented in the training N-grams. TA-Prompting's risk of hallucinated captions depends on the LLM backbone quality.

Practical implications

For embodied AI systems, integrating discriminative prompting into video understanding pipelines would significantly improve retrieval and moment identification. The temporal extension strategy provides a fallback when within-clip discrimination fails, though it trades precision for uniqueness. Combining CDP's prompt selection mechanism with TA-Prompting's temporal anchoring could provide both content-level and temporal-level discriminability for egocentric video understanding.

Theme 2: Evidence for 3D Scene Understanding as a Core Capability for Egocentric Video

Why this theme matters

The grounded note identifies that "no clear method for ingesting 3D information into VLMs yet" remains an open challenge. Yet 3D scene understanding—maintaining awareness of objects not currently visible, reasoning about spatial relationships, understanding reachability and occlusion—is fundamental to embodied intelligence. A system that cannot track objects outside the camera's field of view cannot anticipate, plan, or reason about physical manipulation effectively.

What the literature says

OSNOM provides the most validated pipeline for 3D object tracking. The Lift-Match-Keep approach achieves 57% correct localization after 120 seconds out of view, compared to 33% for other 3D methods and 17% for 2D tracking. The key insight is that interaction context—objects near hands are likely being

manipulated—significantly improves matching accuracy because manipulated objects maintain spatiotemporal continuity.

Ego3DT provides a complementary zero-shot approach using open-vocabulary detection (GLEE) and 3D reconstruction (DUSt3R). This approach achieves 1.04x to 2.90x HOTA improvements compared to prior methods without task-specific training, making it more broadly deployable.

The geometry-guided camera motion paper provides evidence that motion-related information is inadequately captured during VLM pre-training. Probing experiments reveal that VideoLLMs struggle to distinguish object motion from camera motion, which is a prerequisite for accurate 3D scene understanding from egocentric video.

How strong the support is

The support for 3D tracking itself is strong—OSNOM and Ego3DT both demonstrate validated approaches with strong quantitative results. However, the support for integrating 3D understanding into VLMs is weak. The geometry-guided paper identifies the gap but does not demonstrate end-to-end integration. The grounded note observation that "no clear method for ingesting 3D information into VLMs yet" remains accurate.

What does not transfer cleanly

OSNOM requires SLAM/ARIA sensor data that may not be available in all egocentric datasets. Ego3DT relies on pre-trained depth estimation that may struggle with textureless or reflective surfaces. Neither approach has been demonstrated in an integrated VLM pipeline that can query 3D scene state through natural language.

Practical implications

For embodied AI, 3D tracking provides concrete capabilities (object localization, occlusion reasoning, reachability queries) but integration with VLMs remains an open challenge. Systems may need to maintain separate 3D tracking modules and VLM reasoning modules rather than unified architectures.

Theme 3: Diffusion Models as the Dominant Approach for Hand Trajectory Forecasting

Why this theme matters

Hand pose forecasting in egocentric video requires handling stochastic future motion, accounting for camera ego-motion, and predicting positions when hands are not visible. The grounded note specifically highlights the "Invisible Ego Hand" challenge—forecasting hand positions outside the camera's field of view—as an important capability for anticipatory systems.

What the literature says

Three papers provide validated diffusion-based approaches for hand trajectory prediction:

EgoH4 uses diffusion-based prediction with body pose constraints to enable forecasting when hands leave the frame. The method jointly denoises hand and body joints, with body joints serving as kinematic constraints on hand motion. A visibility predictor estimates hand visibility, improving capability for dealing with invisible hands. Results show +3.4cm improvement on hand trajectory ADE and +5.1cm on hand pose MPJPE.

MADiff introduces Motion-Driven Selective Scan (MDSS) that integrates camera ego-motion into the denoising process. The approach disentangles hand motion from camera motion with temporal causality, achieving state-of-the-art across five public datasets with real-time inference capability.

MMTwin uses twin latent diffusion models with separate models for camera ego-motion prediction and hand trajectory prediction, coordinated through a hybrid Mamba-Transformer module.

How strong the support is

Strong and well-validated. All three approaches demonstrate quantitative improvements over baselines. EgoH4 specifically addresses the out-of-sight forecasting challenge, MADiff demonstrates real-time capability, and MMTwin provides architectural insights about separating camera and hand motion prediction.

What does not transfer cleanly

EgoH4 requires accurate body pose estimation as input and is evaluated on manual hand joints rather than tool interaction. MADiff requires camera pose estimation. MMTwin requires multiple input modalities and adds coordination complexity. None of the approaches have been demonstrated in integrated embodied AI systems that use hand forecasting for manipulation planning.

Practical implications

Diffusion-based approaches are the clear winner for hand trajectory forecasting. For embodied AI, integrating these methods would enable anticipatory manipulation planning—predicting what hands will do next to prepare tools, clear workspace, or avoid collisions. The core capability gap is that all three approaches are validated on manual hand joints; performance on tool-holding hands (utensils, appliances, objects) has not been established. This matters because embodied AI manipulation primarily concerns tool use, not bare hand movement.

Theme 4: VLM Failures on Physical Understanding Are Fundamental, Not Merely Performance Gaps

Why this theme matters

The grounded note documents that VLMs perform near-random on object motion questions in the HD Epic VQA benchmark. More critically, the note observes that "stronger VLMs improve plausibility but can actually hide grounding failures." This distinction matters enormously: a model that scores near-random reveals its failure modes, while a model that generates confident but incorrect answers masks them.

What the literature says

The MVP Benchmark provides the most systematic evidence. Using minimal-change pairs (visually similar video pairs with identical questions but opposing answers), the benchmark forces models to demonstrate genuine physical understanding rather than exploiting shortcuts. Results show a 52.7 percentage point gap between human performance (92.9%) and the best open-source Video-LLM (40.2%).

Crucially, models that perform well on standard benchmarks perform poorly on MVP, confirming that standard benchmarks allow shortcut exploitation. The MVP finding validates the concern from the grounded note: VLMs may appear capable on standard benchmarks while remaining fundamentally limited in genuine physical understanding.

The geometry-guided camera motion paper provides a mechanistic explanation: probing experiments reveal that motion-related information is inadequately captured during pre-training. This is not a parameter count or architecture issue—it is a fundamental gap in what information the models learn to represent.

HD-EPIC VQA results confirm the pattern: even powerful models like Gemini Pro achieve only 37.6% accuracy on the 26.6K question benchmark, with particularly poor performance on fine-grained action understanding, 3D perception, and object motion questions.

How strong the support is

Very strong and well-controlled. The MVP benchmark design (minimal-change pairs) specifically controls for shortcut exploitation. The findings are consistent across multiple benchmarks and model families. The mechanistic explanation from geometry-guided probing provides theoretical support for why the gap exists.

What does not transfer cleanly

The MVP benchmark is limited to multiple-choice format and may not cover all physical reasoning types. The geometry-guided approach requires additional supervision during training, which may not be feasible for existing VLMs without retraining. Solutions for preventing shortcuts in VLMs are not yet demonstrated.

Practical implications

Current VLMs cannot serve as reliable foundations for embodied AI systems requiring physical understanding. This is not a temporary performance gap but a fundamental limitation of how VLMs represent motion, space, and causality. Systems may need to maintain separate physical reasoning modules rather than relying on VLMs for all understanding.

Theme 5: Scene-Conditioned Generation as a Complement to Discriminative Understanding

Why this theme matters

The grounded note discusses two complementary directions: discriminative understanding (captioning what is happening) and generative instruction (showing how to do something). ShowHowTo addresses the generative direction, generating step-by-step visual instructions from an input image and text instructions. The grounded note records that this approach achieves better human preference ratings than original videos in some cases—a striking finding that suggests generated instructions can exceed real-world demonstrations in quality.

What the literature says

ShowHowTo uses automatically collected training data from 1 million instructional videos, resulting in 0.6M sequences of image-text pairs. The video diffusion model generates image sequences conditioned on input scene and text instructions. A scene consistency mechanism maintains object presence and scene layout from input across generated steps.

StableWorld provides complementary evidence about maintaining consistency in long video generation. The dynamic frame eviction mechanism—monitoring geometric consistency and evicting degraded frames—addresses the error accumulation problem that causes generated frames to deviate from initial state.

How strong the support is

Moderate. ShowHowTo demonstrates state-of-the-art on three evaluation dimensions, and the human preference finding is striking. StableWorld provides quantitative evidence of improved temporal coherence. However, both approaches are limited to the cooking domain, and ShowHowTo's generated images may have artifacts.

What does not transfer cleanly

Current methods are cooking-specific and may not generalize to other domains without significant retraining. The grounded note identifies state consistency errors when common priors contradict action sequences—a challenge that StableWorld addresses but does not fully solve.

Practical implications

For embodied AI, scene-conditioned generation could enable task demonstration, tutorial generation, and action visualization. However, current methods are domain-limited and artifact-prone, requiring significant development before reliable deployment.

5. Integrated Assessment for the Current Project

The research survey reveals a clear picture of where egocentric video understanding stands: individual component capabilities are maturing rapidly, but integration into unified systems remains the critical bottleneck.

Directions that look most justified

Discriminative prompting for unique captioning is the most mature and validated approach. CDP demonstrates 33 percentage point improvements in R@1 on egocentric benchmarks, is captioner-agnostic, and scales with base model quality. For a project requiring video understanding in repetitive action domains, integrating discriminative prompting would provide immediate and substantial benefits.

3D object tracking is well-validated for maintaining object awareness outside the camera's field of view. OSNOM achieves 57% localization after 120 seconds, and the interaction context insight (objects near hands maintain spatiotemporal continuity) provides a practical mechanism for improving matching accuracy. Ego3DT's zero-shot approach extends this capability to arbitrary object classes without task-specific training.

Diffusion-based hand trajectory forecasting has achieved strong results across multiple papers and datasets. EgoH4's body pose constraints enable out-of-sight forecasting, MADiff's motion disentanglement handles camera ego-motion, and MMTwin's twin model architecture provides a modular approach. For systems requiring anticipatory manipulation understanding, these methods are ready for integration.

Assumptions that are weakly supported

The assumption that 3D information can be ingested into VLMs is weakly supported. The geometry-guided paper identifies the gap but does not demonstrate integration. The grounded note's observation that "no clear method for ingesting 3D information into VLMs yet" remains accurate. Systems should plan for separate 3D tracking and VLM reasoning modules rather than unified architectures.

The assumption that VLMs will improve to handle physical understanding is also weakly supported. The MVP benchmark demonstrates that even powerful VLMs exploit shortcuts rather than developing genuine physical reasoning. Geometry-guided training is promising but requires additional supervision that may not be feasible for existing models.

Branches that deserve further testing

End-to-end integration of discriminative captioning, 3D tracking, and hand forecasting into unified pipelines has not been demonstrated. Even though individual components are validated, their interaction effects, error propagation, and latency characteristics in integrated systems remain unknown.

Shortcut-resistant evaluation for embodied AI tasks deserves systematic testing. The MVP benchmark design (minimal-change pairs) could be applied to egocentric video benchmarks to expose whether VLMs are genuinely understanding physical interactions or exploiting shortcuts.

6. Unresolved Questions and Decision-Critical Gaps

Gap 1: No demonstrated approach for 3D-to-VLM integration

The grounded note identifies that "no clear method for ingesting 3D information into VLMs yet" remains an open challenge, and the literature survey confirms this gap persists. The geometry-guided camera motion paper provides a pathway but does not demonstrate end-to-end integration. A decision-critical question for the project is whether to pursue unified architectures (VLM with geometry-guided training) or modular architectures (separate 3D tracking + VLM reasoning).

Gap 2: VLM shortcuts obscure genuine capabilities and failures

The MVP benchmark demonstrates that models performing well on standard benchmarks perform poorly when shortcuts are controlled. For evaluating embodied AI systems, this means standard VQA benchmarks may not reveal where models genuinely fail. The project needs evaluation methodology that exposes shortcut exploitation rather than allowing inflated scores.

Gap 3: Hand forecasting validation is limited to manual hand joints

EgoH4 is evaluated on manual hand joints rather than tool interaction. For embodied AI systems that must understand manipulation (using tools, operating appliances, handling objects), the gap between hand joint forecasting and tool interaction forecasting is significant and needs validation.

Gap 4: Scene-conditioned generation is cooking-domain-limited

ShowHowTo and StableWorld are validated in the cooking domain. Extending scene-conditioned generation to other embodied AI domains (workshop, garage, outdoor) would require dataset collection and model retraining. The scope of this extension is not yet characterized.

Gap 5: Temporal correspondence in video generation remains challenging

The grounded note identifies "continue improving temporal correspondence in video generation" as a suggested next step. StableWorld addresses this through

frame eviction but does not fully solve the problem. For applications requiring high-fidelity generated video (tutorials, demonstrations, simulations), the remaining gaps in temporal correspondence need systematic characterization.

Gap 6: Evaluation benchmarks are domain-specific

HD-EPIC, EPIC-KITCHENS, and Ego4D are all kitchen-focused. Other embodied AI domains (workshop, garage, outdoor activities) lack comparable annotated datasets. Cross-domain generalization of methods validated in kitchen settings is not established.

7. Recommended Next Steps

Action 1: Integrate discriminative prompting into the video understanding pipeline

CDP provides an immediately applicable improvement for unique captioning in repetitive action domains. The approach is captioner-agnostic, computationally tractable (5.8 seconds via CDPNet approximation), and demonstrates larger improvements on better base models. This action matters because caption ambiguity directly impairs retrieval, moment identification, and text-based video understanding. The implementation path is clear: integrate the prompt bank and CDPNet into the existing captioning pipeline.

Action 2: Deploy 3D object tracking as a separate module

OSNOM and Ego3DT provide validated approaches for 3D object tracking. Given the lack of 3D-to-VLM integration methods, deploy these as separate modules that maintain world-coordinate object memory and expose queries through a structured interface. This approach trades architectural elegance for immediate capability. Track objects in 3D space using depth estimation and SLAM, maintain object identity using interaction context (hand proximity), and expose spatial reasoning queries (visibility, occlusion, reachability) through a separate reasoning module rather than attempting VLM integration.

Action 3: Validate hand forecasting on tool interaction scenarios

EgoH4, MADiff, and MMTwin demonstrate hand trajectory forecasting on manual hand joints. Before relying on these methods for embodied AI manipulation understanding, validate their performance on tool interaction scenarios (holding utensils, operating appliances, manipulating objects). This action matters because manipulation understanding in embodied AI primarily concerns tool use rather than bare hand movement. Compare diffusion-based forecasting accuracy with and without tool contact, and characterize failure modes.

Action 4: Design shortcut-resistant evaluation for embodied AI benchmarks

Apply the MVP minimal-change pair methodology to create embodied AI-specific benchmarks that control for shortcut exploitation. Generate video pairs that are visually similar but differ in a critical physical interaction detail, and require models to answer questions about both correctly. This action matters because standard benchmarks may be giving inflated impressions of VLM physical understanding. The evaluation design should follow the MVP methodology but target embodied AI tasks: object manipulation, tool use, spatial reasoning.

Action 5: Characterize geometry-guided training feasibility

Investigate whether geometry-guided training approaches can be applied to existing VLMs without full retraining. The geometry-guided camera motion paper identifies that motion information is inadequately captured during pre-training, but the training approach (additional geometric supervision) may be applicable as fine-tuning rather than full retraining. This action matters because it determines whether existing VLMs can be improved with targeted training or require architectural changes.

Action 6: Investigate StableWorld frame eviction for consistency maintenance

StableWorld's dynamic frame eviction mechanism addresses error accumulation in video generation by monitoring geometric consistency and evicting degraded frames. Investigate whether this mechanism can be adapted for scene-conditioned instruction generation to address the state consistency errors (raw vs cooked) documented in the grounded note. This action matters because consistency errors undermine the reliability of generated instructions.

8. Key Risks, Caveats, and Evidence Boundaries

Risk 1: Individual component validation does not imply integrated system performance

The literature demonstrates strong results for discriminative captioning, 3D tracking, hand forecasting, and VLM evaluation in isolation. However, no literature demonstrates end-to-end systems combining these components. Integration may introduce error propagation, latency issues, and interaction effects not visible in component-level evaluation. This is an evidence-transfer risk: the strength of evidence for individual components does not transfer to integrated systems.

Risk 2: VLM physical understanding failures are fundamental, not solvable by scale

The MVP benchmark (52.7 percentage point gap) and HD-EPIC results (37.6% Gemini Pro) demonstrate that VLM failures on physical understanding are not performance gaps that can be closed by larger models or more training. Geometry-guided probing confirms that motion information is inadequately captured during pre-training. This is a methodological risk: current VLMs may be

architecturally limited for physical reasoning, requiring fundamental changes rather than incremental improvements.

Risk 3: 3D tracking quality depends on depth estimation

OSNOM achieves 57% localization accuracy but requires depth estimation that has known limitations. Ego3DT relies on DUSt3R for 3D reconstruction. In environments with textureless surfaces, reflective objects, or poor lighting, depth estimation quality degrades. This is a data-quality risk: 3D tracking performance is bounded by depth estimation quality, which varies across environments.

Risk 4: Hand forecasting validation gap for tool interaction

EgoH4 is validated on manual hand joints, not tool interaction. For embodied AI systems that primarily involve tool use, this represents a significant evidence boundary. Forecasting accuracy for hands holding tools may differ substantially from bare hand forecasting due to kinematic constraints, occlusion, and object-hand attachment. This is an evaluation risk: the evidence for hand forecasting does not directly transfer to the embodied AI use case of tool interaction understanding.

Risk 5: Scene-conditioned generation is cooking-domain-limited

ShowHowTo and StableWorld are validated in the cooking domain. Extending to other domains requires new datasets, retraining, and validation. The grounded note specifically concerns egocentric video understanding, which spans many domains beyond cooking. This is a domain-transfer risk: the evidence for scene-conditioned generation does not generalize across embodied AI domains.

Risk 6: Shortcut-resistant evaluation may reveal worse VLM performance than standard benchmarks suggest

The MVP benchmark demonstrates that models performing well on standard benchmarks perform poorly when shortcuts are controlled. Applying this methodology to embodied AI evaluation may reveal that VLM capabilities are substantially worse than standard benchmarks indicate. This is an evaluation risk: deploying VLMs based on standard benchmark performance may lead to systems that fail in real-world scenarios where shortcuts are unavailable.

Risk 7: HD-EPIC is validation-only, limiting training data availability

The HD-EPIC dataset is designed as validation only and should not be used for fine-tuning. This constraint limits the availability of high-quality 3D annotations for training. For projects requiring training on egocentric video understanding, this represents a data-quality risk: lower-quality annotations may be necessary for training, potentially limiting model performance.

Risk 8: Dense object tracking annotation is expensive

The grounded note identifies that "dense object tracking (frame-by-frame bounding boxes) is expensive to annotate." Scaling 3D object tracking to new domains requires substantial annotation investment. This is a methodological risk: the current evidence for 3D tracking comes from extensively annotated datasets that may not be feasible to replicate for novel domains.