VIMA: Video Intelligence for Multimodal Annotation

RGB-Only Spatial Evidence for Construction Video

Philip Chen Joshua Lin Stephen Hung Lucas He
Hacktech 2026 Spatial Intelligence Track

Abstract

Egocentric construction video is a poor substrate for direct visual-language-model (VLM) judgment. Frontier VLMs can describe scenes fluently while hallucinating objects, mis-estimating distance, or making progress claims without recoverable evidence. We present VIMA, an RGB-only spatial evidence pipeline for hardhat video. VIMA decomposes video understanding into specialist perception stages—object proposals, segmentation masks, monocular depth, depth-delta frame filtering, reconstruction, and episodic memory—before any VLM writes an answer. The core contribution is a depth-delta pre-filter: consecutive frames whose normalized depth disagreement exceeds $\delta = 0.25$ are treated as over-motion pairs and withheld from reconstruction. In hackathon pilot runs, this removed 57% of unstable construction frame pairs, reduced a masonry-site translation norm by 70% (0.370 to 0.111), and produced an object-event memory that links answers to frames, masks, depth maps, and candidate episodes. VIMA reframes construction AI from "ask a VLM what happened" to "verify spatial claims against evidence."

1 Introduction

Construction progress and compliance are spatial problems. A useful system must answer not only *what* appears in a frame, but also *where* it is, whether it persists across time, which physical object changed, and whether the claim can be audited later. Egocentric-video benchmarks such as Ego4D show why first-person video needs memory and retrieval rather than isolated image captioning [2]. Raw hardhat footage is even less forgiving: fisheye distortion, motion blur, repeated materials, partial occlusion, scaffold geometry, and multiple workers make many single-frame judgments ambiguous.

Our first pass was intentionally simple: give frontier VLMs construction frames and ask for safety, distance, or progress judgments. That was not reliable enough. The model outputs sounded confident, but the spatial grounding was brittle: objects were hallucinated, distance estimates drifted, and activity labels were generated without evidence frames. These are not harmless captioning errors. In a construction workflow, an unsupported spatial claim can credit the wrong work, miss a hazard, or create an audit trail that cannot be defended.

VIMA therefore treats the VLM as a final synthesizer, not as the spatial sensor. This follows the broader lesson of spatial VLM work: language models can benefit from explicit spatial structure, but the structure should be externalized rather than assumed [1]. The pipeline builds a structured memory from RGB video first, then lets an answer layer retrieve and explain from that memory.

Figure 1: The basic VIMA evidence unit: a raw hardhat frame next to a depth-conditioned view. The depth map is not used as metric truth; it is used as an RGB-derived spatial signal for ordering, stability checks, and reviewer-facing evidence.

2 Target Failure

The task we target is **spatially verifiable work evidence**: given hardhat video, determine whether an activity claim is supported by frames, object tracks, depth/order evidence, and reconstruction diagnostics. Humans can inspect the footage and say, for example, "masonry activity is visible near this block wall at these timestamps." A VLM can often produce that sentence, but it does not naturally return the evidence chain that makes the sentence inspectable.

The point is not that VLMs are useless. They are useful summarizers. The failure is asking one model to perform detection, temporal binding, depth reasoning, action attribution, and written explanation in a single opaque step. VIMA separates those responsibilities.

Attempt	Observed failure
Raw VLM object judgment	Hallucinated site objects in frame-level reasoning.
Raw VLM distance judgment	Scaffold distance estimates were off by large factors in pilot checks.
Single-frame activity label	Over-labeled activity when temporal context was removed.
VLM-only answer	Could answer fluently without citing frames, tracks, or geometry.

Table 1: Why VIMA moved away from direct VLM judgment. These pilot observations define the failure mode: unsupported spatial text is not enough for jobsite evidence.

3 Pipeline

VIMA is an RGB-only evidence stack. It does not assume LiDAR, IMU, a site model, or private construction labels. The input is a video; the output is an object-event memory whose claims point back to concrete visual evidence.

3.1 Frame-Level Evidence

Frames are sampled from video and converted into three aligned evidence views: raw RGB, object/mask overlays, and depth-conditioned overlays. Figure 2 shows the reviewer-facing version of that representation. The raw view preserves visual context; masks localize entities such as workers, walls, guardrails, scaffold, and material stacks; depth adds near/mid/far ordering to those entities.

3.2 Specialist Models Before Language

VIMA uses a specialist-first stack:

1. **Object proposals**: detector labels for workers, scaffold, concrete block walls, guardrails, material stacks, tools, and open edges.
2. **Masks and tracks**: box-prompt segmentation converts rectangles into object support regions and maintains identities across nearby frames.
3. **Monocular depth**: DepthAnythingV2-style relative depth estimates provide near/mid/far ordering for each object region [7].
4. **Geometry**: MASt3R/COLMAP-style reconstruction estimates camera motion and sparse structure on filtered frame sets [3, 4, 5].
5. **Memory**: object facts are grouped into episodes that can be retrieved for a user query.

The key design choice is that each intermediate output remains visible. The system does not collapse detections into a caption and throw away the evidence. If the final answer says a wall-related work candidate exists, the dashboard can still show the raw frame, the object region, the depth view, and the episode that produced the claim.

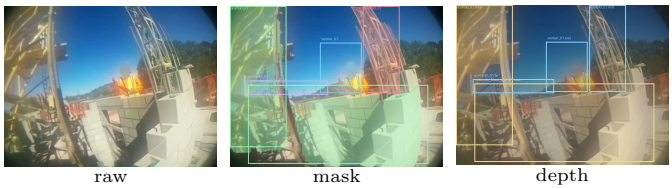

Figure 2: Aligned evidence views for one construction frame. The same frame can be inspected as raw video, segmentation/mask evidence, or depth-conditioned spatial evidence.

4 Depth-Delta Filtering

Construction video contains many adjacent frames that are adjacent in time but poor for geometry: fast head turns, blur, partial occlusion, and worker-body motion. We found that using every frame pair made reconstruction unstable, so VIMA uses depth disagreement as a simple RGB-only spatial filter.

For consecutive frames I_t and I_{t+1}, a monocular depth model produces normalized maps $D_t, D_{t+1} \in [0,1]^{H \times W}$. VIMA computes

$$\Delta_D(t) = \sqrt{\frac{1}{HW} \sum_{x,y} (D_t(x,y) - D_{t+1}(x,y))^2}. \quad (1)$$

If $\Delta_D(t) > \delta$ with $\delta = 0.25$, the pair is marked as over-motion and skipped before reconstruction. This is not claiming the depth model has metric accuracy. It is using relative depth as a stability oracle: when two adjacent frames disagree too much in spatial layout, they are likely bad inputs for pose estimation.

The simplest picture of the contribution is the transition from raw visual evidence to depth-conditioned evidence and then to downstream spatial memory. The filter is cheap, model-agnostic, and requires only RGB video. It also gives an interpretable failure explanation: the system can say a pair was skipped because its estimated spatial layout changed too aggressively, rather than silently feeding every frame into geometry.

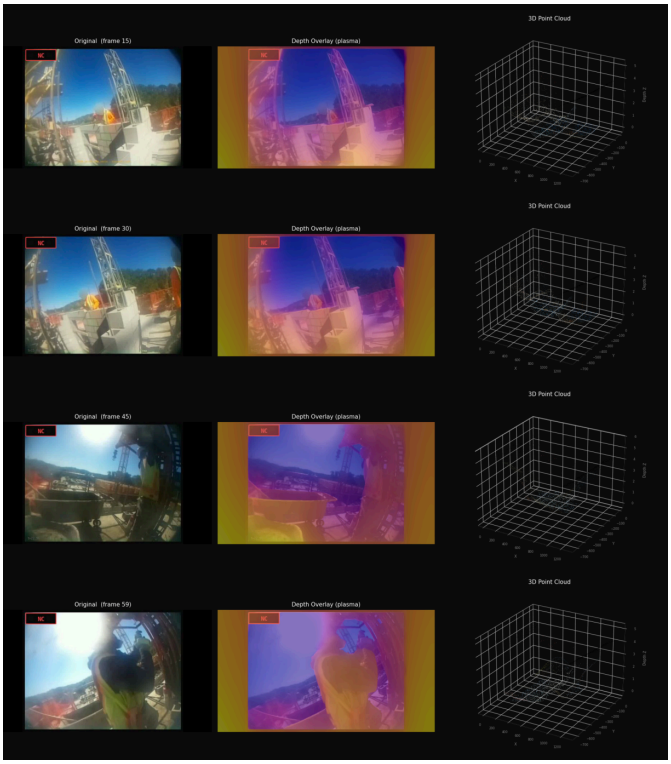

Figure 3: Three-way comparison used in the paper narrative: raw spatial evidence, depth-conditioned reconstruction signal, and downstream risk/memory visualization. Keeping the views side by side makes the pipeline inspectable instead of relying on a single generated caption.

5 Geometry and Reconstruction

After filtering, VIMA runs reconstruction only on the frame pairs that pass the depth-delta check. Figure 4 shows the resulting reconstruction diagnostics: source frames, predicted depth/confidence products, and geometry outputs. This stage is where VIMA converts an egocentric video stream into a spatial memory substrate.

The scalar results in Table 2 should be read as a pilot validation of the mechanism: filtering the wildest egocentric frame pairs improved downstream geometry. The important product implication is that not all frames should become evidence. VIMA separates evidence frames from unstable frames before any reward, safety, or productivity logic consumes them.

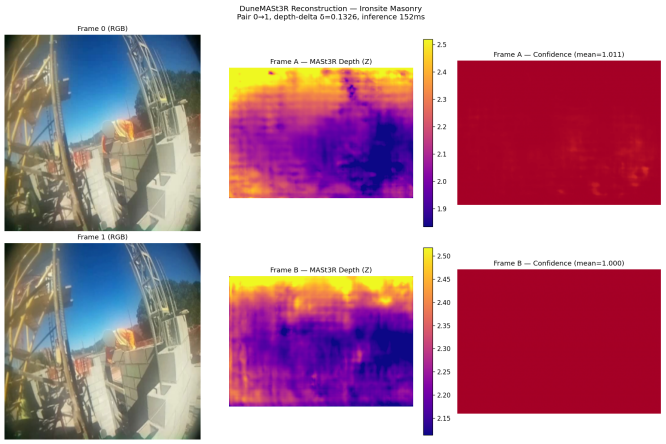

Figure 4: Reconstruction diagnostic panel for the construction clip. The view collects RGB evidence, depth products, and reconstruction confidence so a reviewer can see why a frame pair was accepted or rejected.

Component	Metric	Pilot result
Depth-delta filter	Dropped pairs	57%
MASt3R on TUM	RPE_t reduction	59%
Masonry-site clip	Translation norm	0.370 → 0.111
Masonry-site clip	Translation improvement	70%
Masonry-site clip	Confidence change	+32%
COLMAP site run	Registered images	19/31
COLMAP site run	Sparse points	1,770
COLMAP site run	Reprojection error	1.199px

Table 2: Preliminary geometry results from the hackathon pipeline. These are pilot numbers, not a frozen benchmark, but they show that frame selection materially changes reconstruction quality.

6 Object-Event Episodic Memory

The reconstruction layer is useful, but the dashboard does not expose geometry alone. It exposes *episodes*: compact records that bind object tracks, time ranges, spatial relations, confidence, and evidence frames. Example episode types include `masonry_work_candidate`, `safety_edge_context`, `scaffold_zone_visible`, `foreground_worker_present`, and `material_staging_visible`.

This memory layer is the main difference between VIMA and a video-captioning demo. A caption says "a worker is near a block wall." An episode stores the frame ids, object ids, mask support, depth ordering, confidence, and time span that make the statement checkable. When a user asks whether masonry work occurred near a wall, VIMA retrieves relevant episodes and then asks the language layer to summarize only from those records.

Artifact	Value	Role
Sampled frames	10	Demo subset
Frame events	35	Object/depth facts
Episodes	10	Retrieval units
Top confidence	0.875	Masonry candidate
Qwen local answer	Success	Open VLM probe
Gemini answer	Success, quota-limited	Synthesis probe

Table 3: Object-event memory artifacts used by the dashboard and answer layer.

7 Answer Layer

VIMA supports both API VLMs and local open-weight probes. Gemini produced cleaner natural-language synthesis from the memory artifact when quota was available. Qwen2-VL-2B ran locally and confirmed that retrieved evidence frames and memory records can be consumed without a remote API, though it followed citation instructions less reliably [6].

The important constraint is that the answer layer is evidence-bound. For the query *"Was there masonry work happening near the wall?"*, the answer references retrieved episodes and frame names instead of making a free-form judgment from pixels alone. A typical answer can say: *yes, there is candidate masonry work near a concrete block wall, supported by episode ids, evidence frames, worker/wall co-presence, and proximity where available.* This keeps the VLM useful without pretending it is a calibrated spatial measurement system.

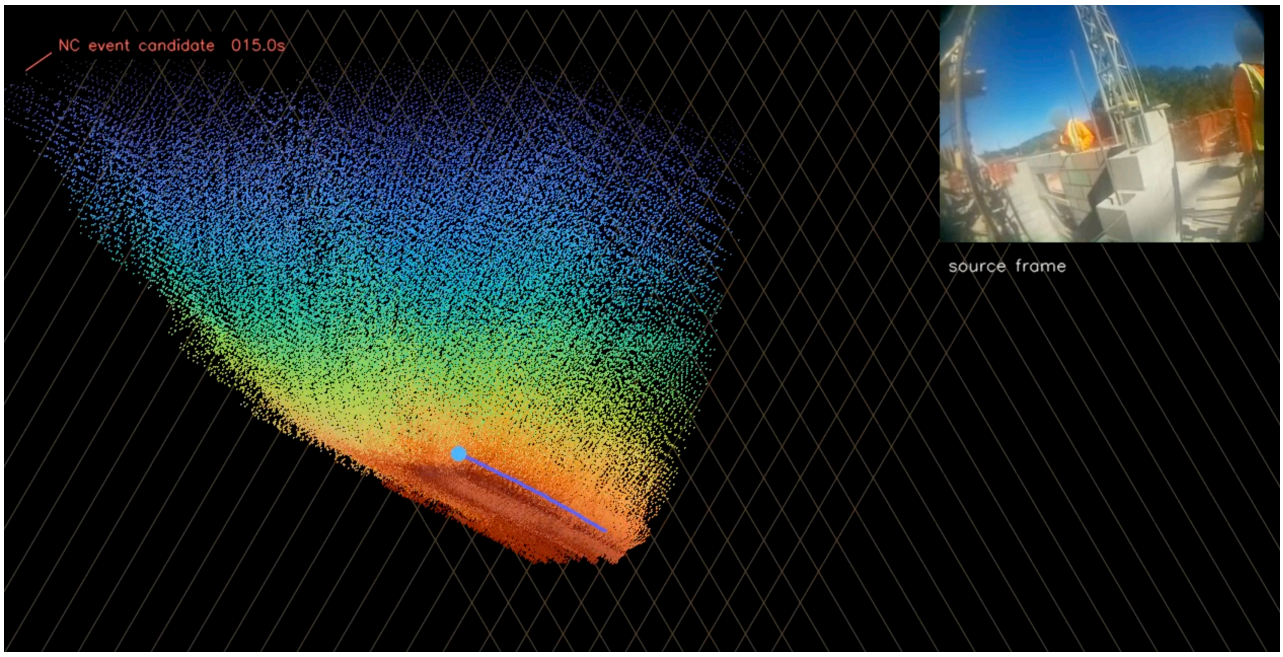

Figure 5: Depth overlays and point-cloud views across multiple frames. These views show how the same raw video can be converted into spatial evidence frames rather than treated as a flat slideshow.

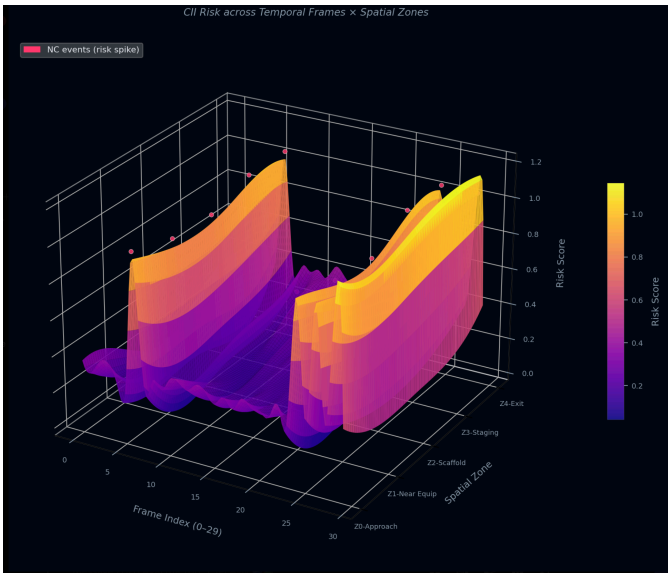

Figure 6: Zone-aware memory visualization. VIMA's claim is not only timestamp-aware; it tries to attach work evidence to spatial zones so downstream review can ask where the evidence occurred.

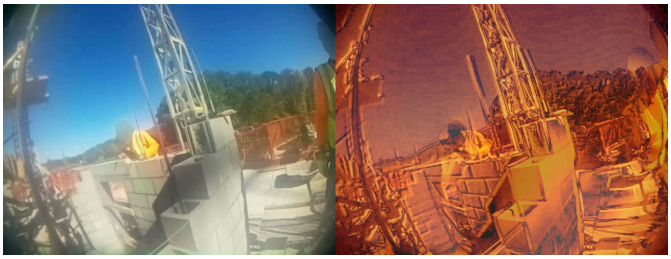

Figure 7: A compact A/B view of raw frame and depth-transformed frame. This visualization is useful in demos because it makes the RGB-only spatial signal visible without asking the judge to inspect arrays or logs.

8 Dashboard

The dashboard is an internal evidence-review console rather than a consumer rewards app. It shows the current claim, raw/mask/depth evidence views, run statistics, retrieved answers, and an object-event timeline. Figure 7 gives a compact A/B view of the RGB frame and depth-transformed frame; Figure 8 shows how frame-level evidence becomes zone-aware temporal evidence.

The dashboard is intentionally conservative. It does not present the VLM answer as ground truth. It presents a claim, a confidence, and the artifacts needed to inspect it. This distinction matters for construction deployment: the system should help reviewers find evidence faster, not replace evidence with fluent text.

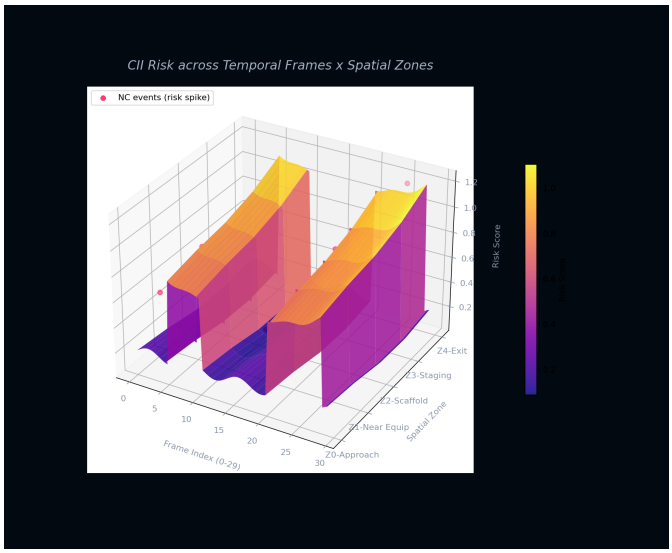

Figure 8: Temporal risk surface over frame index and spatial zone. This is the downstream form of the memory: object-event episodes become spatially indexed evidence that can support review, auditing, or risk triage.

9 Limitations

The current system is a hackathon prototype. The reported numbers come from small pilot runs and should be treated as directional until repeated on a frozen dataset. DepthAnythingV2 is relative, not metric, so VIMA uses it for ordering, filtering, and stability checks rather than absolute measurements. The CII classifier runs were inconsistent: one full run over-labeled frames as non-contributory, while smaller temporal batches produced more plausible P/C/NC splits. This argues for temporal batching and human-reviewed benchmark labels before any safety-critical deployment.

There are also limits to the evidence layer itself. Masks can fail on blurry frames, object proposals can miss small tools, and depth can be distorted by fisheye imagery or reflective surfaces. VIMA's defense is not that every intermediate model is correct. Its defense is inspectability: every claim is decomposed into visible artifacts that a reviewer can accept, reject, or relabel.

10 Conclusion

VIMA's central claim is that construction VLM systems need a spatial evidence layer before language synthesis. Direct VLM prompting is too brittle for jobsite claims because it can hallucinate objects, mis-estimate distance, and omit the evidence path. By combining RGB-derived depth, depth-delta frame filtering, object masks, reconstruction, and episodic memory, VIMA turns hardhat video into auditable spatial claims. The most promising technical contribution is the simplest one: a monocular-depth disagreement check that removes unstable frame

pairs before geometry and improves downstream reconstruction quality in pilot runs.

References

[1] Boyuan Chen, Zhuo Xu, Sean Kirmani, et al. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. *arXiv preprint arXiv:2401.12168*, 2024.

[2] Kristen Grauman, Andrew Westbury, Eugene Byrne, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[3] Vincent Leroy, Yohann Cabon, and Jerome Revaud. Grounding image matching in 3d with mast3r. *arXiv preprint arXiv:2406.09756*, 2024.

[4] Johannes L. Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[5] Johannes L. Schönberger, Enliang Zheng, Marc Pollefeys, and Jan-Michael Frahm. Pixelwise view selection for unstructured multi-view stereo. In *European Conference on Computer Vision*, 2016.

[6] Peng Wang, Shuai Bai, Sinan Tan, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

[7] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *arXiv preprint arXiv:2406.09414*, 2024.