Technical Summary: COARSE CORRESPONDENCES Boost Spatial-Temporal Reasoning in Multimodal Language Models

Part I: Problem Formulation, Methods, and Evidence

Kai-Yu Lu

2025/10/5

1 Research Problem and Motivation

Multimodal Language Models (MLLMs) are increasingly deployed in real environments where agents must understand three-dimensional (3D) space and temporal dynamics from image sequences or multi-view observations. Although modern MLLMs excel in linguistic and visual-linguistic tasks, prior evaluations indicate pronounced weaknesses in 3D spatial reasoning and long video understanding. The central problem addressed in this study is how to strengthen spatial-temporal reasoning in off-the-shelf MLLMs while keeping the solution simple, lightweight, training-free, and architecture-agnostic.

Motivation Existing approaches often require specialized 3D inputs, model modifications, or costly task-specific fine-tuning. The work summarized here proposes a visual prompting strategy that exploits *coarse* object correspondences across frames or views. The method aims to unlock 3D and temporal reasoning ability that general MLLMs can already approximate if key spatial correspondences are made explicit in the visual input.

2 Related Work

Prior lines include: providing 3D data or point clouds as inputs; designing specialized architectures for 3D tasks; and performing supervised fine-tuning with 3D data for spatial or temporal understanding. For long videos, research also explores video-first encoders or Socratic pipelines that convert frames to text before prompting a text-only LLM. Visual prompting has improved grounding on single images, such as Set-of-Mark or axis labels. The present work differs by jointly enhancing spatial and temporal reasoning through instance-level correspondences over image sequences, without architectural changes or task-specific training.

3 Dataset Construction

The study evaluates on established benchmarks rather than constructing a new dataset. The core datasets are summarized in Table 1.

Dataset	Description and Task Focus	Split Used
ScanQA	3D scene question answering requiring recognition, 3D localization, and embodied capability from multiple views of indoor scenes.	Validation
OpenEQA (EM-EQA)	Open-vocabulary episodic memory questions that stress spatial understanding across time while an agent moves through a 3D environment.	EM-EQA subset
EgoSchema	Very long-form egocentric videos with multiple-choice questions that require broad temporal context.	Validation subset
R2R (VLN-CE)	Vision-and-Language Navigation in continuous environments focusing on spatial-temporal grounding for action decisions.	val-unseen
SQA3D	Out-of-domain 3D QA for generalization analysis on open-source models.	Evaluation

Table 1: Benchmarks used for evaluation.

4 Query Protocol and Task Definitions

Input: a natural language question Q and a sequence or a set of images $[I_1, \ldots, I_n]$ that depict either a video over time or multiple viewpoints of a static scene.

Output: an answer $\hat{A}$ generated by an MLLM.

Goal: design a visual prompt $P(\cdot)$ to transform the image set into prompted images $[I'_1, \ldots, I'_m]$ so that a general MLLM $M(\cdot)$ can infer 3D spatial relations and temporal dependencies more reliably and sample-efficiently.

5 Modeling Approach

Terminology

Multimodal Language Model (MLLM): a large language model augmented with a vision encoder to accept images and text.
Coarse Correspondences (CC): instance-level associations of the same physical object across frames or views, represented as simple overlaid marks rather than dense pixel-level correspondences.
Tracklets: temporally consistent instance identities produced by an off-the-shelf tracker across a sparsified set of frames.

Method overview

The pipeline comprises four steps: tracking correspondences, frame sparsification, selection of prominent tracklets, and visualization of marks on images. The prompted images are then fed to a general MLLM during inference.

Formalization

Prompted inference Given Q and images $[I_1, \ldots, I_n]$, the prompting function $P(\cdot)$ produces a reduced and annotated set $[I'_1, \ldots, I'_m]$ used by an MLLM $M(\cdot)$:

$$[I'_1, \ldots, I'_m] \;=\; P([I_1, \ldots, I_n]), \quad \hat{A} \;=\; M\big([I'_1, \ldots, I'_m],\, Q\big). \tag{1}$$

Symbols: I_i denotes the i-th original image, I'_j denotes a prompted image after overlay, $m \ll n$ is the retained frame count, and $\hat{A}$ is the model output. **Intuition**: P encodes cross-frame object identity and location as visual hints so that M can resolve spatial perspective and temporal continuity.

Tracking and sparsification A class-agnostic video object tracker generates an instance mask $M_i \in \mathbb{N}^{H\times W}$ for each original image I_i with integer instance identifiers per pixel. To control cost, frames are uniformly downsampled to indices $s_1, \ldots, s_m$ with corresponding masks M_{s_j}.

Selecting prominent correspondences Overlaid marks must be informative but not occlusive. Instances are scored by occurrence frequency and cumulative area across the retained m masks. For an instance identifier ID:

$$\mathrm{Freq}(\mathrm{ID}) \;=\; \sum_{i=1}^{m} \mathbf{1}\{\mathrm{ID} \in M_{s_i}\}, \qquad \mathrm{Area}(\mathrm{ID}) \;=\; \sum_{i=1}^{m} \sum_{p\in M_{s_i}} \mathbf{1}\{p = \mathrm{ID}\}. \tag{2}$$

Symbols: $\mathbf{1}\{\cdot\}$ is the indicator function, p ranges over pixel positions in a mask, ID is an instance identity. **Use**: instances are sorted by descending Freq and then by Area as tiebreaker. The top-k tracklets $T_1, \ldots, T_k$ are retained.

Visualizing correspondences For each retained tracklet T_i that appears in mask M_{s_j}, a small marker with label T_i is placed at the instance centroid on image I_{s_j}. The centroid is:

$$(\bar{x}_{ij}, \bar{y}_{ij}) \;=\; \frac{\sum_{(x,y)} (x,y)\, \mathbf{1}\{M_{s_j}(x,y) = T_i\}}{\sum_{(x,y)} \mathbf{1}\{M_{s_j}(x,y) = T_i\}}. \tag{3}$$

Symbols: (x, y) index pixel coordinates, $(\bar{x}_{ij}, \bar{y}_{ij})$ is the centroid of instance T_i in frame s_j. **Interpretation**: the centroid gives a stable placement for a small, legible marker that communicates identity continuity without masking important content.

Design choices and ablations

The study examines the number, size, and type of marks. Too many or too large marks degrade accuracy by occluding content. Adding segmentation contours can help grounding, while filling masks harms visibility. Selecting a handful of high-frequency instances with moderate-size circular markers strikes an effective balance.

6 Empirical Results

Evaluation protocol and metrics

For ScanQA, standard caption-style metrics are used, including BLEU, METEOR, ROUGE-L, and CIDEr. For OpenEQA EM-EQA, the evaluation reports an average matching score of generated answers against

reference using a strong judge model. For EgoSchema, multiple-choice accuracy is measured on validation questions. For navigation on R2R val-unseen, Success Rate (SR), Oracle Success Rate (OS), Success weighted by Path Length (SPL), Trajectory Length (TL), and Navigation Error (NE) are used.

Closed-source MLLMs at inference time

With GPT-4V and GPT-4O, Coarse Correspondences improves spatial 3D QA and long video understanding while using fewer frames. On ScanQA, adding CC to GPT-4O yields notable gains across BLEU-2, METEOR, ROUGE-L, and CIDEr, and surpasses 3D-specific models that rely on fine-tuning or 3D point clouds. On OpenEQA EM-EQA, CC improves GPT-4V and GPT-4O accuracy with fewer views. On EgoSchema, CC improves GPT-4O by a sizeable margin using only eight uniformly sampled frames from three-minute videos, outperforming many fine-tuned approaches that process far more frames. On R2R navigation, CC increases SR and SPL and reduces NE, revealing an explicit benefit for embodied tasks that depend on robust spatial-temporal grounding.

Open-source MLLMs with training and inference

For open models such as LLaVA, the study first instruction-tunes on mixed data to interpret marks and accept multiple images, then applies CC in inference. On ScanQA, LLaVA with CC improves over the same model trained without CC. Zero-shot generalization to SQA3D also improves. Interestingly, applying CC only during training yields benefits at test time even without CC prompts, indicating a useful data augmentation effect.

Camera motion invariance

A diagnostic benchmark with forward and reversed scan orders shows that plain GPT-4O accuracy drops under reversed ordering, while CC restores invariance and raises the harmonic mean accuracy substantially. This indicates improved robustness to camera motion and viewpoint sweep direction.

7 Summary

This work introduces a training-free, architecture-agnostic visual prompting technique that overlays a small number of instance-level correspondence markers at object centroids across sparsified frames or views. The method systematically enhances 3D spatial and temporal reasoning in general MLLMs across diverse benchmarks, reduces the number of required frames, and improves embodied navigation metrics. For open-source models, CC can be used in training to deliver gains at inference and to improve out-of-domain generalization. The analysis clarifies effective design regimes for mark count, size, and representation, and shows improved invariance to camera motion.

Technical Details

Notation and symbols

- Q: natural language question.
- $[I_1, \ldots, I_n]$: original images from a video or multi-view scene.
- $[I'_1, \ldots, I'_m]$: prompted images after overlay, $m \ll n$.

- $M(\cdot)$: a general-purpose MLLM.
- $M_i \in \mathbb{N}^{H \times W}$: instance mask for I_i with integer identifiers per pixel.
- $T_1, \ldots, T_k$: selected top-k instance tracklets by frequency and area.
- $(\bar{x}_{ij}, \bar{y}_{ij})$: centroid of instance T_i in frame s_j.

Implementation choices

- **Tracker**: a class-agnostic video object tracker that produces consistent instance IDs across frames.
- **Frame sampling**: uniform downsampling to control token cost for MLLMs.
- **Overlay rendering**: small circular markers with numeric labels at instance centroids; optional contours to aid grounding; avoid opaque fills to reduce occlusion.

Evaluation metrics

- **ScanQA**: BLEU, METEOR, ROUGE-L, CIDEr.
- **OpenEQA EM-EQA**: average answer match score using a strong evaluator.
- **EgoSchema**: multiple-choice accuracy on long videos.
- **R2R**: SR, OS, SPL, TL, NE for navigation quality.

Baselines

- **3D-specific**: models tailored for 3D QA with specialized architectures or 3D inputs.
- **General MLLMs**: GPT-4V, GPT-4O, Gemini, Claude; with and without CC.
- **Open-source**: LLaVA; with and without CC during training and inference.

Limitations and Future Directions

Limitations. The approach relies on tracker quality and consistent instance IDs. Over-marking can occlude salient content and reduce performance. The current design uses instance-level rather than dense point correspondences, which may miss fine-grained motion cues or small objects.

Future directions. Potential extensions include: learning to predict the most useful instances to mark; adapting mark placement to object scale and task context; combining coarse marks with sparse contours that preserve visibility; exploring automatic frame selection policies by uncertainty or entropy; and using CC as a curriculum or augmentation for broader multimodal pretraining.