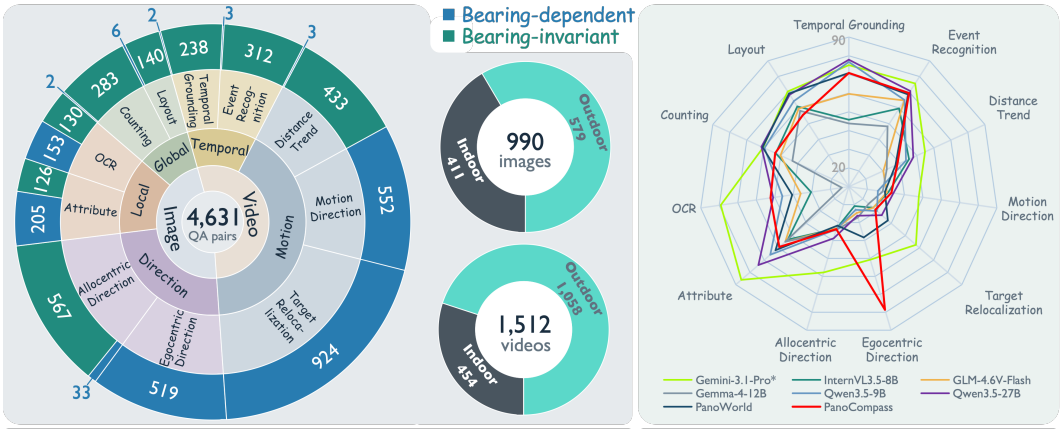
PANOARENA: ZERO-SHOT BENCHMARK FOR PANORAMIC UNDERSTANDING

Anonymous authors
Paper under double-blind review



(a) Examples of PanoArena.



(b) Data distribution. (c) Performance.

Figure 1: **Overview of PanoArena.** Blue and teal denote camera bearing-dependent and bearing-invariant questions, respectively. (a) Representative QA examples illustrate static understanding of panoramic images and dynamic understanding of videos. (b) Task hierarchy, question counts, and indoor/outdoor scene distributions. (c) Evaluation results on PanoArena with ERP input. "*" denotes thinking mode.

ABSTRACT

Existing evaluations of panoramic understanding remain largely image-centric and fail to systematically assess panorama-specific properties, particularly view-dependent spatial reasoning under the complete 360° field of view (FoV). To bridge this gap, we introduce **PanoArena**, a comprehensive benchmark with 4,631 samples for evaluating general-purpose panoramic understanding across both images and videos. PanoArena adopts a unified question-answering format

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.

with more than 10 sub-tasks spanning semantic and spatial understanding, where questions are further categorized into *bearing-dependent* and *bearing-invariant* settings based on whether the observer's view direction is required for reasoning. We systematically benchmark existing multimodal large language models (MLLMs) across different panoramic representations under zero-shot settings. We further propose PanoCompass, a training-free method, enabling a cleaner assessment of intrinsic panoramic reasoning rather than task-specific memorization. Extensive experiments reveal substantial limitations of current MLLMs, particularly in bearing-dependent spatial reasoning, providing new insights and directions for general-purpose panoramic understanding. The benchmark data and relevant code will be made available to the public.

1 INTRODUCTION

The 360° panoramic understanding has become important as panoramas provide a complete field of view (FoV) that captures comprehensive visual information about the surrounding environment (Lin et al., 2025a). This capability has benefited numerous applications, including immersive media (Lin et al., 2025b; Chi et al., 2026), virtual reality (Zhu et al., 2025), embodied intelligence (Zheng et al., 2025; Wan et al., 2025; Liu et al., 2026), and virtual cinematography (Hu et al., 2017; Wang et al., 2025a).

Despite the remarkable progress in panoramic understanding, existing models mainly focus on individual modalities or specific tasks, lacking a unified framework to systematically exploit panoramic information across diverse tasks. Recent multimodal large language models (MLLMs), such as PanoWorld (Wang et al., 2026) and PLM (Fan et al., 2026), have attempted to provide more general-purpose understanding (Zhou et al., 2025; Yang et al., 2026; Tran et al., 2026). However, existing evaluations remain image-centric and fail to assess panorama-specific properties, such as rotation invariance and view bearing-dependency, which are better exposed through temporal variations and spatial understanding.

An intuitive solution is to establish a comprehensive video-level benchmark by drawing upon well-established benchmarks for perspective videos. Nevertheless, panoramic videos differ fundamentally from their perspective counterparts in representation and viewing characteristics, making existing evaluation protocols not directly applicable. For example, a panoramic video can support reasoning about the direction and distance of specific objects relative to an assigned observer, owing to its complete 360° FoV.

To address this issue, we re-formulate the benchmark, naming PanoArena, in a general question-answering (QA) format in both panoramic images and videos with main focus on spatial understanding. The resulting benchmark contains 990 images with 2,164 static questions and 1,512 video clips with 2,467 dynamic questions, for 4,631 questions in total. It covers indoor and outdoor scenes and separates six static and five dynamic sub-tasks. Across 11 sub-tasks, every question records whether its answer depends on the camera bearing, yielding 2,402 bearing-dependent and 2,229 bearing-invariant questions.

After that, we systematically benchmark various existing MLLMs under two zero-shot evaluation settings. On one hand, we explore them across different panoramic representations, including equirectangular projections (ERPs), cubemap projections (CMPs), and collage layouts (COLs). On the other hand, we further propose a training-free method — PanoCompass, without exposing them to any benchmark-specific tasks or questions. Following the spirit of domain adaptation, PanoArena better reflects the model's spatial reasoning ability rather than its capacity to memorize task-specific patterns from the same distribution.

Our experiments reveal four patterns. First, perspective-rich cubemap and collage inputs often improve bearing-dependent spatial reasoning over a single ERP, although the gains are task- and model-dependent. Second, bearing-invariant questions are generally easier than camera-dependent ones, exposing a persistent gap in camera-relative spatial reasoning. Third, increased model scale improves overall performance but does not remove this gap, and thinking mode is not uniformly beneficial across models and modalities. Finally, PanoCompass improves the matched ERP baseline

most strongly on egocentric direction and bearing-dependent questions, while its gains on global layout are mixed.

The main contributions of our work are as follows:

- We introduce PanoArena, a general-domain benchmark spanning panoramic images and videos, with 4,631 QA questions across 11 static and dynamic sub-tasks.
- We define question-level camera-bearing dependence annotations and a controlled evaluation protocol for accuracy, yaw consistency, and all-angle correctness.
- We provide a unified comparison of general-purpose and panorama-oriented MLLMs across ERP, cubemap projection, and collage representations under zero-shot evaluation.
- We propose PanoCompass, a training-free and agentic baseline, and analyze how representation, model scale and reasoning mode affect panoramic understanding.

2 Related Work

2.1 General MLLM Benchmarks

With the rapid growth of multimodal large language models (MLLMs) (Qwen Team, 2026; Niu et al., 2025), the community has introduced a range of benchmarks to assess their perception and reasoning abilities. For holistic evaluation, the image-domain benchmarks MME (Fu et al., 2025a), MMBench (Liu et al., 2024), and SEED-Bench (Li et al., 2024a), together with the knowledge-intensive MMMU (Yue et al., 2024), build multi-dimensional protocols spanning coarse-grained perception to expert-level reasoning; in the video domain, MVBench (Li et al., 2024b), Video-MME (Fu et al., 2025b) and Video-MME-v2 (Fu et al., 2026) cover multi-task video understanding, while TimeLens-Bench (Zhang et al., 2026) and OMTG Bench (Xu et al., 2026b) further endow video LLMs with precise moment localization. For fine-grained perception, V*Bench (Wu & Xie, 2024) and PinpointBench (Gong et al., 2026) expose models' weaknesses on small targets and subtle visual patterns.

However, the visual inputs of these works are almost exclusively ordinary normal field-of-view images and videos and do not cover the challenges brought by 360° panoramas.

2.2 Panoramic MLLM Benchmarks

VQA360° (Chou et al., 2020) pioneers panoramic MLLM evaluation; Pano-AVQA (Yun et al., 2021) extends it to videos, and 360+x (Chen et al., 2024) studies cross-view scene understanding including panoramic videos. Evaluation on panoramic images has since proliferated along complementary axes: omnidirectional spatial reasoning (Dongfang et al., 2026), dense captioning and grounding (Zhou et al., 2025), reasoning-enhanced QA (Zhang et al., 2025b; Chen et al., 2026b), multi-frame cross-view understanding (Zhang et al., 2025a), visual search (Yu et al., 2026), ERP-native spatial reasoning (Wang et al., 2026), 3D questions grounded in dense geometry (Lin & Zheng, 2026), and adverse driving scenes (Fan et al., 2026), while ODI-Bench (Yang et al., 2026) and 360Bench (Tran et al., 2026) cover broad panoramic image QA. Panoramic video evaluation remains far sparser: Fire360 (Tiwari et al., 2025) targets degraded firefighting footage, ORSet (Chen et al., 2026a) formulates omnidirectional referring multi-object tracking, and 360CityArena (Watanabe et al., 2026) builds an embodied navigation benchmark from 360° street videos. Despite this breadth, the same limitations recur across both modalities: the data are built on synthetic scenes (Dongfang et al., 2026; Lin & Zheng, 2026; Chen et al., 2026b; Zhang et al., 2025a) or confined to a single domain (Fan et al., 2026; Tiwari et al., 2025; Chen et al., 2026a; Watanabe et al., 2026), limiting visual diversity; and the dynamic side is served by tracking outputs or sequential navigation (Chen et al., 2026a; Watanabe et al., 2026), with none asking an MLLM whether an object approaches or recedes and in which direction it travels. Orthogonal to all of these, none records whether answering a question requires the camera's reference frame. Our PanoArena fills this gap: over real, in-the-wild 360° images and videos spanning indoor and outdoor scenes, it jointly evaluates panoramic static and dynamic understanding, from basic perception to extensive reasoning, offering the first general-domain MLLM VQA benchmark.

Table 1: **Comparison of PanoArena with existing panoramic MLLM benchmarks.** *Gen.*: unconstrained in-the-wild content rather than a single domain; *Img./Vid.*: evaluating images or videos; *Ego.*: direction expressed in the camera's reference frame; *Allo.*: direction in a reference frame attached to an entity; *Temp.*: associating events with their time spans; *Mot.*: object motion over time (distance trend, motion direction, target relocalization); *Pub.*: publicly available. ✓ = supported, △ = partially covered, blank = not supported.

Benchmark Name	Gen.	Modality		Capability				Pub.
		Img.	Vid.	Ego.	Allo.	Temp.	Mot.	
VQA360° (Chou et al., 2020)		✓		✓				
Pano-AVQA (Yun et al., 2021)	✓		✓	✓				
360+x (Chen et al., 2024)	✓		✓			✓		✓
Fire360 (Tiwari et al., 2025)			✓			✓		✓
OSR-Bench (Dongfang et al., 2026)		✓			✓			✓
Dense360-Bench (Zhou et al., 2025)	✓	✓						
OmniVQA (Zhang et al., 2025b)		✓		✓				
CFPano (Zhang et al., 2025a)		✓		✓				
H*Bench (Yu et al., 2026)	✓	✓		✓				✓
ODI-Bench (Yang et al., 2026)	✓	✓		✓	✓			✓
360Bench (Tran et al., 2026)	✓	✓		✓				✓
PanoEnv (Lin & Zheng, 2026)	✓	✓		✓				✓
PanoVQA (Fan et al., 2026)		✓		✓				✓
PanoSpace-Bench (Wang et al., 2026)	✓	✓		✓	✓			✓
PCSR-Bench (Chen et al., 2026b)		✓		✓	✓			✓
ORSet (Chen et al., 2026a)			✓				△	✓
360CityArena (Watanabe et al., 2026)			✓	✓	✓			✓
PanoArena (Ours)	✓	✓	✓	✓	✓	✓	✓	✓

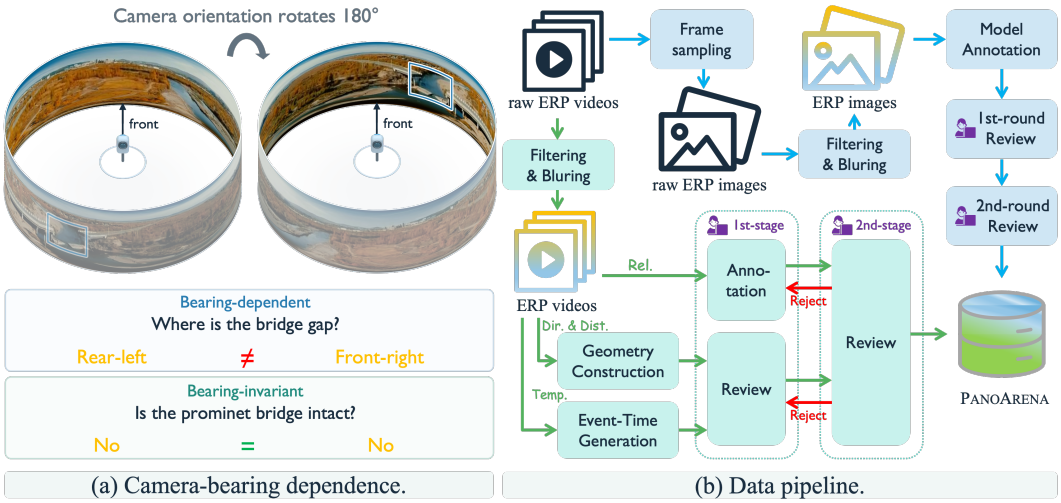
2.3 Panoramic MLLM Methods

Training-free methods utilize the MLLMs for zero-shot or few-shot inference. Omni-CoT (Yang et al., 2026) captions six perspective views and then refines its answer using grounded narrow-FoV crops. Free360 (Tran et al., 2026) assembles a question-relevant scene graph via cubemap view mapping. **Panorama-specific post-training** instead injects spherical priors through data and parameters: 360-R1 (Zhang et al., 2025b), PanoEnv (Lin & Zheng, 2026) and PCSR-Bench (Chen et al., 2026b) apply GRPO with geometry-grounded rewards, Dense360 (Zhou et al., 2025) uses panoramic-aware positional encodings, PanoWorld (Wang et al., 2026) adds spherical spatial cross-attention over a large ERP instruction corpus, and PLM (Fan et al., 2026) uses a panoramic sparse attention module for global information fusion. **Agentic exploration** treats the panorama as an environment: HVS (Yu et al., 2026) formulates humanoid visual search over local views, while EAGLE-360 (Xu et al., 2026a) reasons global-to-local and adapts rolled positional encodings.

3 PanoArena

3.1 Overview

PanoArena evaluates panoramic understanding along two complementary dimensions. Static Understanding is built on 990 equirectangular images and asks 2,164 questions about a single 360° observation; Dynamic Understanding is built on 1,512 panoramic video clips and asks 2,467 questions that additionally require reasoning over time. Both halves cover indoor and outdoor scenes: Static Understanding contains 411 images about indoor scenes and 579 about outdoor scenes, while Dynamic Understanding contains 454 videos about indoor scenes and 1,058 about outdoor scenes. Together, they comprise 4,631 questions across six static and five dynamic sub-tasks.



(a) Camera-bearing dependence. (b) Data pipeline.

Figure 2: **Camera-bearing dependence and dataset construction.** (a) Rotating the camera by 180° changes the bridge gap's camera-relative direction from rear-left to front-right, while the answer about whether the bridge is intact remains unchanged. (b) Both pipelines begin with data filtering and face blurring. The image pipeline uses model-generated annotations followed by two rounds of human review. The video pipeline uses manual target annotation or geometric and temporal proposals, followed by human review and iterative refinement.

3.2 Camera-Bearing Dependence Label

We fix one convention throughout the benchmark: the centre column of the ERP image is the camera's forward direction, the horizontal axis is azimuth, the vertical axis is elevation, and the seam at $\pm 180°$ lies directly behind the camera.

Alongside the task taxonomy, we label every question by whether answering it requires this camera reference frame. As shown in Fig 2(a), a question is **bearing-dependent** on either of two routes: its answer is a bearing in the convention above, or the object it asks about is designated by such a bearing. It is **bearing-invariant** otherwise, in which case both its statement and its answer survive a yaw rotation of the camera unchanged. The label is therefore assigned per question rather than per sub-task, yielding 2,402 bearing-dependent and 2,229 bearing-invariant questions.

3.3 Task Description

3.3.1 Static Understanding

Static understanding evaluates an MLLM's ability to perceive and reason about 360° scenes from a single panoramic image. It consists of three tasks and six sub-tasks.

Direction Reasoning. Because panoramic images capture the surrounding scene in all directions, directional questions can refer to any bearing and use different reference frames. *Egocentric Direction* asks for the approximate direction of an object relative to the camera, which requires mapping an ERP position to a physical bearing. *Allocentric Direction* instead fixes the reference frame on a person or object in the scene, so the model must first recover that entity's facing direction and then re-express the target's bearing in that reference frame.

Local Perception. This task evaluates the model's ability to perceive local visual details within a continuous panoramic scene that usually contains rich information. *Attribute* requires the model to identify attributes (e.g., color, size, or shape) of a designated target object. *OCR* similarly requires the model to locate and read text on a designated target. Both sub-tasks identify the target either by an egocentric bearing or by a semantic description alone; the latter requires searching the whole panoramic scene for the target.

Global Perception. This task requires holistic scene understanding by integrating information across the full 360° field of view. *Counting* requires the model to count objects across the scene. *Layout* assesses the model's understanding of spatial structure, scene topology, and structural relationships through a complete panoramic observation.

3.3.2 Dynamic Understanding

Dynamic understanding evaluates an MLLM's ability to comprehend temporal information and object motion in panoramic videos. It consists of two tasks and five sub-tasks, all formulated as multiple-choice questions.

Temporal Perception. Because a panoramic clip records all directions simultaneously, multiple events can occur at different times and azimuths; the model must locate the relevant event within this broader context. Given a specific event description, *Temporal Grounding* requires the model to pinpoint the approximate temporal window in which it occurs. Conversely, *Event Recognition* requires the model to identify the events occurring within a given temporal interval.

Motion Reasoning. Panoramic videos record objects in all directions around the camera. This task evaluates the model's ability to track these objects and reason about their motion over time. *Distance Trend* requires the model to infer how the distance between a target and the camera, or between two targets, evolves over a given window, choosing among several trends: closer, farther, first closer then farther, first farther then closer, and unchanged. *Motion Direction* asks for a target's local direction of movement around a specified time. We express its horizontal displacement in eight 45° sectors: front, front-right, right, rear-right, back, rear-left, left, and front-left. The camera's forward ($+z$) and rightward ($+x$) axes define these directions and are shared by all targets. For example, an object on the camera's left that moves along $+z$ has a frontward motion direction. The labels describe displacement independently of the target's current bearing or facing direction; stationary targets are labeled as not moving. *Target Relocalization* presents a reference observation of a target together with the full video and asks for the target's bearing at another specified timestamp. The model must identify the same object at the target timestamp and determine its camera-relative bearing, making this the dynamic counterpart of Egocentric Direction.

3.4 Data Pipeline

Appendix A details the PanoArena construction procedures. As shown in Fig 2(b), we construct our dataset through separate image and video pipelines, combining automatic candidate generation with task-specific human verification. We consider model automatic labeling as an auxiliary tool to screen a rough range for human annotation.

For images, we sample one frame per source video and use Qwen3.5-27B (Qwen Team, 2026) to check ERP validity and visual informativeness. SAM3 (Carion et al., 2026) are used to segment human face and we apply local Gaussian blurring to the retained panoramas. Each retained ERP image is paired with six labeled cubemap views, providing both global context and local detail. Gemini-3.1-Pro-Preview (Google, 2026) then proposes questions across the six static sub-tasks using this composite and manually annotated examples.

For motion reasoning, we select dynamic, single-shot ERP clips, including segments extracted from longer videos. We apply the same face-blurring procedure to the retained ERP videos on a per-frame basis. On clips with limited camera motion, SAM3 (Carion et al., 2026) object tracks and DAP (Lin et al., 2026) depth estimates provide geometric proposals for distance trends and motion directions. Target relocalization uses human-confirmed object identities at two timestamps, with answers computed from the target masks' panoramic bearings. For temporal perception, Qwen3.5-27B (Qwen Team, 2026) generates event descriptions with temporal boundaries from paired perspective-rendered clips; these segments yield event-recognition and temporal-grounding questions evaluated on the corresponding panoramas.

Human annotation and verification refine the candidate data and remove ambiguous items. For all video tasks, first-stage annotation or QA refinement is followed by second-stage review, which can return items to the first stage for revision and another review. A subset of second-stage items is independently assessed by two reviewers. The final video test set contains 2,467 samples from 1,512 clips. Across all logged paired second-stage reviews, including samples not retained in the

final set, 3,012 paired review records yield 78.1% content agreement. Appendix A also summarizes how many times each retained sample underwent second-stage review. Multiple-choice options are reordered to reduce answer-position bias.

4 PANOCOMPASS

To explore whether the inherent agentic ability of a foundation MLLM (Qwen Team, 2026) can decompose complex panoramic questions into manageable observations, we propose PanoCompass as a training-free agentic baseline for panoramic images and videos. It connects question-guided evidence acquisition and answer generation through a common tool interface. The same pretrained MLLM performs visual localization, observation planning, and evidence reading, while the tools provide spatial views, temporal observations, or geometric measurements. The original ERP is the source for all observations and remains available for direct answering. Each tool preserves the coordinate or timestamp metadata needed to interpret its output in the source panorama. The tools are follows.

Projection. For spatial detail in an image, the model specifies a view center on the ERP and a field of view between 20° and 120°. The tool converts the center into yaw and pitch and renders up to four perspective views, recording their orientation in the original camera frame. The answer reader receives these views together with any reference image provided by the question.

Temporal cropping. For a time-specific video query, the tool constructs a primary observation covering the queried interval or a two-second neighborhood of the queried timestamp, shifted within the video at its boundaries. A timestamp query also receives an independently decoded frame near that time. If the video reader identifies missing evidence, it can request one reselection from the original ERP. The model can request up to three supplementary clips to provide additional context.

Calculation. For camera-relative direction questions, the model localizes the target directly on the original ERP. Let (u_c, v_c) be the predicted box center in an ERP of width W and height H, measured from its top-left corner. The tool computes azimuth and elevation in radians as

$$\theta = \frac{2\pi u_c}{W} - \pi, \qquad \phi = \frac{\pi}{2} - \frac{\pi v_c}{H}. \tag{1}$$

Zero azimuth points forward, positive azimuth points right, and positive elevation points upward. The resulting direction is matched to the answer option with the smallest angular separation, completing the geometric answer deterministically.

The process returns the latest valid answer, retaining the first answer if reselection fails. Empty or invalid final predictions are scored as incorrect.

5 EXPERIMENTS

5.1 METRICS

Each question is scored as either correct or incorrect, and overall model performance is measured by **accuracy**. For multiple-choice questions, we perform an exact match between the ground-truth and the model's response. For counting questions, we likewise require an exact match between the model's response and the ground truth. For OCR questions, we normalize both the ground truth and the prediction by removing whitespace and punctuation and converting all characters to lowercase, and then perform an exact match; a response is regarded as correct only when the two strings are identical.

5.2 EVALUATION SETTINGS

We evaluate closed-source commercial model Gemini-3.1-Pro-Preview (Google, 2026), open-source general models like InternVL3.5 (Wang et al., 2025b), GLM-4.6V-Flash (Hong et al., 2025), Gemma-4 (Gemma Team, 2026), Qwen3.5 (Qwen Team, 2026) and open-source panorama-optimized methods including PanoWorld (Wang et al., 2026) and our PanoCompass. For each model, where supported, we evaluate three visual inputs: ERP, cubemap projection (CMP), and

Table 2: **Main results on PanoArena.** *In.*: Visual input format. We report accuracy (%) on each sub-task, grouped into Static Understanding (panoramic images) and Dynamic Understanding (panoramic videos). *Ego./Allo.*: egocentric/allocentric direction; *Attr.*: attribute; *OCR*: scene-text recognition; *Cnt.*: counting; *Lay.*: layout; *Grd./Rec.*: temporal grounding/event recognition; *Dist.*: distance trend; *Dir.*: motion direction; *Rel.*: target relocalization; *Dep./Inv.*: bearing-dependent/invariant; *All*: overall accuracy. "*" denotes thinking mode. Within the closed-source and open-source general model groups separately, **bold** and underline mark the best and second-best results in each column.

Model	In.	Static Understanding						Dynamic Understanding					Summary		
		Direction		Local		Global		Temporal		Motion			Bearing		All
		Ego.	Allo.	Attr.	OCR	Cnt.	Lay.	Grd.	Rec.	Dist.	Dir.	Rel.	Dep.	Inv.	
Closed-source commercial model															
Gemini-3.1-Pro*	ERP	50.7	57.8	86.1	79.2	60.4	70.5	75.0	75.6	**54.8**	47.8	**57.5**	58.2	64.9	61.5
	CMP	95.2	71.0	92.1	85.2	58.9	78.8	**76.2**	**82.2**	52.8	**51.1**	46.5	65.7	**69.8**	**67.7**
	COL	**95.6**	**72.5**	**93.1**	**92.6**	**63.2**	**79.5**	64.6	71.1	47.9	47.5	51.1	**67.2**	67.5	67.4
Open-source general models															
InternVL3.5-8B	ERP	20.8	33.3	53.2	30.4	53.0	61.0	45.8	59.7	45.4	31.3	27.2	28.6	46.8	37.4
	CMP	16.8	36.5	62.2	33.2	48.4	59.6	40.0	51.1	40.1	30.1	24.5	27.6	44.5	35.7
	COL	42.8	36.0	67.1	41.7	55.1	55.5	40.0	55.2	41.7	28.4	27.1	35.1	46.3	40.5
InternVL3.5-8B*	ERP	19.3	25.0	43.2	31.1	38.9	38.4	47.9	57.8	43.1	30.1	26.3	26.8	40.3	33.3
	CMP	41.4	31.8	63.4	30.0	33.0	51.4	40.4	50.5	41.3	29.9	24.9	32.9	40.8	36.7
	COL	34.9	30.0	63.1	39.6	36.5	47.9	31.7	49.8	38.8	27.9	27.2	32.9	39.2	35.9
GLM-4.6V-Flash	ERP	24.7	33.2	55.0	36.0	50.2	59.6	59.6	64.8	30.3	30.6	27.2	30.1	45.7	37.6
	COL	51.6	35.7	73.4	52.7	53.7	67.8	57.5	62.2	30.7	32.2	26.4	39.4	48.0	43.5
GLM-4.6V-Flash*	ERP	26.8	32.0	59.2	40.3	52.3	61.0	58.3	63.8	41.7	33.0	**33.7**	34.2	48.2	40.9
	COL	67.4	35.3	77.0	56.2	53.0	62.3	59.2	59.4	39.4	34.4	30.2	45.8	48.9	47.2
Gemma-4-12B	ERP	27.2	34.2	55.0	13.8	43.2	58.2	43.8	48.3	32.3	28.3	24.0	27.7	39.7	33.5
	CMP	30.1	33.8	67.1	39.9	46.7	68.5	44.6	52.1	28.9	33.3	22.9	31.9	42.8	37.1
	COL	56.1	35.2	69.5	13.8	38.2	56.2	31.2	45.4	30.5	30.4	25.3	36.7	37.4	37.0
Qwen3.5-4B	ERP	25.6	31.2	58.6	41.0	49.8	63.7	71.2	69.2	39.0	24.3	28.7	29.7	49.8	39.4
	CMP	20.6	31.3	62.8	40.6	50.2	61.6	51.2	57.8	38.1	24.6	27.1	29.2	45.1	36.9
	COL	45.3	32.8	77.0	61.1	55.4	63.7	72.9	65.4	40.6	22.8	31.3	38.1	52.4	45.0
Qwen3.5-4B*	ERP	27.2	35.2	59.8	39.6	49.8	57.5	57.9	62.9	35.1	29.7	20.7	28.9	46.7	37.4
	CMP	**80.5**	**42.0**	79.8	44.5	47.0	61.6	39.6	33.0	13.5	15.2	3.1	33.1	38.5	35.7
	COL	44.5	34.0	76.7	56.9	46.3	49.3	58.3	66.0	29.4	17.8	15.7	31.5	45.6	38.3
Qwen3.5-9B	ERP	22.9	33.8	65.6	45.9	59.6	64.4	76.7	65.1	43.6	25.7	29.9	30.9	53.3	41.7
	CMP	11.0	33.8	74.9	49.8	47.4	66.4	49.6	59.0	39.9	29.2	24.7	28.9	47.3	37.8
	COL	50.3	36.2	84.3	67.8	56.5	63.7	75.0	63.8	42.9	25.5	29.2	40.4	54.3	47.1
Qwen3.5-27B	ERP	26.2	38.8	74.0	50.9	61.4	68.5	**77.9**	**70.8**	**47.9**	34.1	33.3	36.3	57.2	46.4
	CMP	28.1	37.7	84.3	50.9	56.1	**74.0**	64.2	63.2	43.1	**34.6**	25.5	35.2	53.2	43.9
	COL	64.4	39.0	**89.4**	**71.4**	**61.8**	71.2	72.1	70.2	47.0	31.0	32.0	**46.5**	**58.1**	**52.1**
Open-source panorama-optimized methods															
PanoWorld	ERP	38.3	31.7	61.9	40.6	60.7	69.2	70.8	68.6	37.4	29.5	37.6	37.7	51.0	44.1
PanoCompass	ERP	78.6	33.7	59.2	52.3	53.3	55.5	70.8	69.5	38.5	33.0	28.8	44.3	50.7	47.3
Δ vs. Qwen3.5-4B	ERP	+53.0	+2.5	+0.6	+11.3	+3.5	−8.2	−0.4	+0.3	−0.5	+8.7	+0.1	+14.5	+0.9	+8.0

an ERP–cubemap collage (COL). For our PanoCompass, we use Qwen3.5-4B as base model. For reproducibility, the temperature of all experiment is set to 0.

5.3 Main Results

Table 2 summarizes performance across the static and dynamic sub-tasks, together with the bearing-dependent and bearing-invariant subsets.

Overall performance. Gemini-3.1-Pro (Google, 2026) achieves the highest overall accuracy of 67.7% with CMP input, while Qwen3.5-27B (Qwen Team, 2026) achieves the best result among open-source general models at 52.1% with COL input. Motion reasoning remains challenging: the best accuracies across all configurations are 54.8% for distance trend, 51.1% for motion direction, and 57.5% for target relocalization. Every evaluated configuration also scores lower on bearing-dependent questions than on bearing-invariant questions, highlighting the value of reporting camera-bearing dependence alongside overall accuracy.

Table 3: **Camera-yaw invariance on bearing-invariant samples of PanoArena.** Results are aggregated over image and video questions in the bearing-invariant subset using ERP inputs. "*" denotes thinking mode.

Model	Accuracy				Consistency	C-Acc.
	0°	90°	180°	270°		
InternVL3.5-8B	46.8	46.7	47.7	47.7	51.5	30.2
Qwen3.5-4B	49.8	50.6	50.2	50.3	54.2	33.9
Qwen3.5-4B*	46.7	47.2	44.3	46.3	27.1	21.3
Qwen3.5-9B	53.3	54.4	54.1	54.6	55.9	37.2
Qwen3.5-27B	57.2	57.0	56.8	57.4	54.6	**40.2**
PanoWorld	51.0	51.3	50.9	50.6	**60.5**	37.5
PanoCompass	50.7	51.7	51.6	51.6	55.3	34.8

Effect of input representation. COL improves overall accuracy over ERP in every paired configuration, while the effect of CMP varies across models. The scores vary by sub-task: switching Qwen3.5-4B from ERP to COL improves egocentric direction from 25.6% to 45.3%, while motion direction decreases from 24.3% to 22.8%.

Effect of thinking mode. Thinking mode has model-dependent effects. The mixed results indicate that the benefit of additional reasoning depends on the model and input configuration. In many cases, thinking can become trapped in a loop, leading to an inability to converge to a reasonable answer.

Performance of PanoCompass. Using Qwen3.5-4B without training, PanoCompass reaches 47.3% overall accuracy, an improvement of 8.0 percentage points over its ERP baseline, and exceeds the same backbone's COL result of 45.0%. The largest sub-task gains are in egocentric direction, OCR, and motion direction. The improvement is uneven across tasks: layout accuracy decreases from 63.7% to 55.5%, while distance trend and temporal grounding remain close to their ERP baselines. Thus, the tool-assisted baseline substantially improves several directional and local-perception tasks, with further progress needed in global scene understanding and broader temporal reasoning.

5.4 Camera-Yaw Invariance

We evaluate whether model predictions remain stable under camera-yaw rotations that preserve the question's meaning and correct answer. We use the bearing-invariant samples, which account for both the question and choices. Each ERP input is horizontally circularly shifted to represent yaw offsets of $0°$, $90°$, $180°$, and $270°$.

Table 3 reports six metrics on the bearing-invariant subset: accuracy at each of the four angles, consistency, and consist accuracy (C-Acc.). Consistency is the proportion of questions with identical valid predictions across all four angles, regardless of correctness. Consist accuracy is the proportion answered correctly at all four angles.

The non-thinking models show relatively small accuracy variation across yaw angles. PanoWorld has the highest consistency (60.5%), whereas Qwen3.5-27B has the highest C-Acc. (40.2%). The Qwen3.5-4B thinking run is less stable, with consistency of 27.1%.

6 Conclusion

We present PanoArena, a general-domain benchmark with $4{,}631$ questions over panoramic images and videos, covering semantic, spatial, temporal, and motion understanding while explicitly distinguishing bearing-dependent from bearing-invariant reasoning. Our experiments show that panoramic representation, model scale, and thinking mode all affect performance, with camera-relative reasoning and motion remaining challenging. The training-free PanoCompass baseline improves the matched ERP baseline on several tasks, especially egocentric direction, while controlled yaw experiments reveal that prediction stability is distinct from accuracy and is not guaranteed by longer reasoning. We hope PanoArena and its evaluation protocol provide a useful basis for developing more reliable panoramic MLLMs.

AI USE STATEMENT

In this work, we used generative AI tools for polishing paper writing, retrieval related work and coding assistance. We have not used generative AI tools to draft sections of the paper, and generating synthetic datasets or proving mathematical claims is not applicable to this work. Additionally, we used generative AI tools for data preprocessing. We have reviewed all AI-assisted work. All data preprocessed by AI must undergo at least two manual reviews. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris Coll-Vinent, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, Jie Lei, Tengyu Ma, Baishan Guo, Arpit Kalla, Markus Marks, Joseph Greer, Meng Wang, Peize Sun, Roman Rädle, Triantafyllos Afouras, Effrosyni Mavroudi, Katherine Xu, Tsung-Han Wu, Yu Zhou, Liliane Momeni, Rishi Hazra, Shuangrui Ding, Sagar Vaze, Francois Porcher, Feng Li, Siyuan Li, Aishwarya Kamath, Ho Kei Cheng, Piotr Dollar, Nikhila Ravi, Kate Saenko, Pengchuan Zhang, and Christoph Feichtenhofer. SAM 3: Segment anything with concepts. In *International Conference on Learning Representations*, 2026. 6, 14, 15

Hao Chen, Yuqi Hou, Chenyuan Qu, Irene Testini, Xiaohan Hong, and Jianbo Jiao. 360+x: A panoptic multi-modal scene understanding dataset. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 3, 4

Sijia Chen, Zihan Zhou, Yanqiu Yu, En Yu, and Wenbing Tao. ORMOT: A dataset and framework for omnidirectional referring multi-object tracking. *arXiv preprint arXiv:2603.05384*, 2026a. 3, 4

Yuangong Chen, Wai Keung Wong, Jiaxing Li, Ioannis Patras, and Xu Zheng. Beyond localization: A comprehensive diagnosis of perspective-conditioned spatial reasoning in mllms from omnidirectional images. *arXiv preprint arXiv:2605.12413*, 2026b. 3, 4

Jiapeng Chi, Gerd Bruder, Carsten Neumann, Carolina Cruz-Neira, and Dirk Reiners. A multi-layer system for ultra-high-resolution static 360-degree telepresence. *arXiv preprint arXiv:2608.05570*, 2026. 2

Shih-Han Chou, Wei-Lun Chao, Wei-Sheng Lai, Min Sun, and Ming-Hsuan Yang. Visual question answering on 360° images. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, 2020. 3, 4

Zihao Dongfang, Xu Zheng, Ziqiao Weng, Yuanhuiyi Lyu, Danda Pani Paudel, Luc Van Gool, Kailun Yang, and Xuming Hu. Are multimodal large language models ready for omnidirectional spatial reasoning? In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Findings*, 2026. 3, 4

Weijia Fan, Ruiping Liu, Jiale Wei, Yufan Chen, Junwei Zheng, Zichao Zeng, Jiaming Zhang, Qiufu Li, Linlin Shen, and Rainer Stiefelhagen. More than the sum: Panorama-language models for adverse omni-scenes. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026. 2, 3, 4

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, Rongrong Ji, Caifeng Shan, and Ran He. MME: A comprehensive evaluation benchmark for multimodal large language models. In *Advances in Neural Information Processing Systems*, 2025a. 3

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, Peixian Chen, Yanwei Li, Shaohui Lin, Sirui Zhao, Ke Li, Tong Xu, Xiawu Zheng, Enhong Chen, Caifeng Shan, Ran He, and Xing Sun. Video-MME: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025b. 3

Chaoyou Fu, Haozhi Yuan, Yuhao Dong, Yi-Fan Zhang, Yunhang Shen, Xiaoxing Hu, Xueying Li, Jinsen Su, Chengwu Long, Xiaoyao Xie, Yongkang Xie, Xiawu Zheng, Xue Yang, Haoyu Cao, Yunsheng Wu, Ziwei Liu, Xing Sun, Caifeng Shan, and Ran He. Video-MME-v2: Towards the next stage in benchmarks for comprehensive video understanding. *arXiv preprint arXiv:2604.05015*, 2026. 3

Gemma Team. Gemma 4 technical report. *arXiv preprint arXiv:2607.02770*, 2026. 7

Dengxian Gong, Yuanzheng Wu, Haobo Yuan, Zhengdong Hu, Tao Zhang, Yikang Zhou, Shihao Chen, Quanzhu Niu, Kai Wang, Jason Li, Haochen Wang, Lu Qi, Shunping Ji, and Ming-Hsuan Yang. PixelEyes: Decoupling perception and reasoning for pinpoint visual evidence seeking. *arXiv preprint arXiv:2607.00115*, 2026. 3

Google. Gemini 3.1 Pro Preview, 2026. URL https://ai.google.dev/gemini-api/docs/models/gemini-3.1-pro-preview. 6, 7, 8, 14

Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, Shuaiqi Duan, Weihan Wang, Yan Wang, Yean Cheng, Zehai He, Zhe Su, Zhen Yang, Ziyang Pan, Aohan Zeng, Baoxu Wang, Bin Chen, Boyan Shi, Changyu Pang, Chenhui Zhang, Da Yin, Fan Yang, Guoqing Chen, Jiazheng Xu, Jiale Zhu, Jiali Chen, Jing Chen, Jinhao Chen, Jinghao Lin, Jinjiang Wang, Junjie Chen, Leqi Lei, Letian Gong, Leyi Pan, Mingdao Liu, Mingde Xu, Mingzhi Zhang, Qinkai Zheng, Sheng Yang, Shi Zhong, Shiyu Huang, Shuyuan Zhao, Siyan Xue, Shangqin Tu, Shengbiao Meng, Tianshu Zhang, Tianwei Luo, Tianxiang Hao, Tianyu Tong, Wenkai Li, Wei Jia, Xiao Liu, Xiaohan Zhang, Xin Lyu, Xinyue Fan, Xuancheng Huang, Yanling Wang, Yadong Xue, Yanfeng Wang, Yanzi Wang, Yifan An, Yifan Du, Yiming Shi, Yiheng Huang, Yilin Niu, Yuan Wang, Yuanchang Yue, Yuchen Li, Yutao Zhang, Yuting Wang, Yu Wang, Yuxuan Zhang, Zhao Xue, Zhenyu Hou, Zhengxiao Du, Zihan Wang, Peng Zhang, Debing Liu, Bin Xu, Juanzi Li, Minlie Huang, Yuxiao Dong, and Jie Tang. Glm-4.5v and glm-4.1v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. *arXiv preprint arXiv:2507.01006*, 2025. 7

Hou-Ning Hu, Yen-Chen Lin, Ming-Yu Liu, Hsien-Tzu Cheng, Yung-Ju Chang, and Min Sun. Deep 360 pilot: Learning a deep agent for piloting through 360 sports videos. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 2

Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. SEED-Bench: Benchmarking multimodal large language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024a. 3

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, Limin Wang, and Yu Qiao. MVBench: A comprehensive multi-modal video understanding benchmark. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024b. 3

Xin Lin, Xian Ge, Dizhe Zhang, Zhaoliang Wan, Xianshun Wang, Xiangtai Li, Wenjie Jiang, Bo Du, Dacheng Tao, Ming-Hsuan Yang, and Lu Qi. One flight over the gap: A survey from perspective to panoramic vision. *arXiv preprint arXiv:2509.04444*, 2025a. 2

Xin Lin, Shi Luo, Xiaojun Shan, Xiaoyu Zhou, Chao Ren, Lu Qi, Ming-Hsuan Yang, and Nuno Vasconcelos. HQGS: High-quality novel view synthesis with gaussian splatting in degraded scenes. In *International Conference on Learning Representations*, 2025b. 2

Xin Lin, Meixi Song, Dizhe Zhang, Wenxuan Lu, Haodong Li, Bo Du, Ming-Hsuan Yang, Truong Nguyen, and Lu Qi. Depth any panoramas: A foundation model for panoramic depth estimation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026. 6, 17

Zekai Lin and Xu Zheng. Panoenv: Exploring 3d spatial intelligence in panoramic environments with reinforcement learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026. 3, 4

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. MMBench: Is your multi-modal model an all-around player? In *European Conference on Computer Vision*, 2024. 3

Zhongyuang Liu, Min He, Shaonan Yu, Xinhang Xu, Muqing Cao, Jianping Li, Jianfei Yang, and Lihua Xie. Omnivln: Omnidirectional 3d perception and token-efficient llm reasoning for visual-language navigation across air and ground platforms. *arXiv preprint arXiv:2603.17351*, 2026. 2

Quanzhu Niu, Dengxian Gong, Shihao Chen, Tao Zhang, Yikang Zhou, Haobo Yuan, Lu Qi, Xiangtai Li, and Shunping Ji. The 1st solution for 7th lsvos rvos track: SaSaSa2VA. *arXiv preprint arXiv:2509.16972*, 2025. 3

Qwen Team. Qwen3.5: Towards native multimodal agents, 2026. URL https://qwen.ai/blog?id=qwen3.5. 3, 6, 7, 8, 14, 17

Aditi Tiwari, Farzaneh Masoud, Dac Trong Nguyen, Jill Kraft, Heng Ji, and Klara Nahrstedt. Fire360: A benchmark for robust perception and episodic memory in degraded 360° firefighting video. In *Advances in Neural Information Processing Systems*, 2025. 3, 4

Huyen T. T. Tran, Van-Quang Nguyen, Farros Alferro, Kang-Jun Liu, and Takayuki Okatani. 360° image perception with mllms: A comprehensive benchmark and a training-free method. *arXiv preprint arXiv:2603.16179*, 2026. 2, 3, 4

Zhaoliang Wan, Zetong Bi, Zida Zhou, Hao Ren, Yiming Zeng, Yihan Li, Lu Qi, Xu Yang, Ming-Hsuan Yang, and Hui Cheng. RAPID hand: Robust, affordable, perception-integrated, dexterous manipulation platform for embodied intelligence. In *Advances in Neural Information Processing Systems*, 2025. 2

Changpeng Wang, Xin Lin, Junhan Liu, Yuheng Liu, Zhen Wang, Donglian Qi, Yunfeng Yan, and Xi Chen. PanoWorld: Towards spatial supersensing in 360° panorama world. *arXiv preprint arXiv:2605.13169*, 2026. 2, 3, 4, 7

Hai Wang, Xiaoyu Xiang, Weihao Xia, and Jing-Hao Xue. A survey on text-driven 360-degree panorama generation. *IEEE Transactions on Circuits and Systems for Video Technology*, 2025a. 2

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, Zhaokai Wang, Zhe Chen, Hongjie Zhang, Ganlin Yang, Haomin Wang, Qi Wei, Jinhui Yin, Wenhao Li, Erfei Cui, Guanzhou Chen, Zichen Ding, Changyao Tian, Zhenyu Wu, Jingjing Xie, Zehao Li, Bowen Yang, Yuchen Duan, Xuehui Wang, Zhi Hou, Haoran Hao, Tianyi Zhang, Songze Li, Xiangyu Zhao, Haodong Duan, Nianchen Deng, Bin Fu, Yinan He, Yi Wang, Conghui He, Botian Shi, Junjun He, Yingtong Xiong, Han Lv, Lijun Wu, Wenqi Shao, Kaipeng Zhang, Huipeng Deng, Biqing Qi, Jiaye Ge, Qipeng Guo, Wenwei Zhang, Songyang Zhang, Maosong Cao, Junyao Lin, Kexian Tang, Jianfei Gao, Haian Huang, Yuzhe Gu, Chengqi Lyu, Huanze Tang, Rui Wang, Haijun Lv, Wanli Ouyang, Limin Wang, Min Dou, Xizhou Zhu, Tong Lu, Dahua Lin, Jifeng Dai, Weijie Su, Bowen Zhou, Kai Chen, Yu Qiao, Wenhai Wang, and Gen Luo. InternVL3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025b. 7

Kenta Watanabe, Atsuyuki Miyai, Mizuki Takenawa, Kiyoharu Aizawa, and Toshihiko Yamasaki. 360CityArena: A realistic virtual urban navigation benchmark for embodied agents. In *European Conference on Computer Vision*, 2026. 3, 4

Penghao Wu and Saining Xie. V*: Guided visual search as a core mechanism in multimodal llms. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 3

Jingtao Xu, Zizhuo Lin, Jianwen Sun, Yi Yang, and Yawei Luo. Eagle-360: Embodied active global-to-local exploration in 360°. *arXiv preprint arXiv:2607.02479*, 2026a. 4

Qi Xu, Tan Yue, Shihao Chen, Jiahao Meng, Anran Wang, Shunping Ji, Hao Fei, and Xiangtai Li. Towards one-to-many temporal grounding. In *International Conference on Machine Learning*, 2026b. 3

Liu Yang, Huiyu Duan, Ran Tao, Juntao Cheng, Sijing Wu, Yunhao Li, Jing Liu, Xiongkuo Min, and Guangtao Zhai. ODI-bench: Can mllms understand immersive omnidirectional environments? In *International Conference on Learning Representations*, 2026. 2, 3, 4

Heyang Yu, Yinan Han, Xiangyu Zhang, Baiqiao Yin, Bowen Chang, Xiangyu Han, Xinhao Liu, Jing Zhang, Marco Pavone, Chen Feng, et al. Thinking in 360°: Humanoid visual search in the wild. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026. 3, 4

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 3

Heeseung Yun, Youngjae Yu, Wonsuk Yang, Kangil Lee, and Gunhee Kim. Pano-AVQA: Grounded audio-visual question answering on 360° videos. In *IEEE/CVF International Conference on Computer Vision*, 2021. 3, 4

Jun Zhang, Teng Wang, Yuying Ge, Yixiao Ge, Xinhao Li, and Limin Wang. TimeLens: Rethinking video temporal grounding with multimodal llms. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026. 3

Xinshen Zhang, Tongxi Fu, and Xu Zheng. Omnidirectional spatial modeling from correlated panoramas. In *ACM International Conference on Multimedia in Asia*, 2025a. 3, 4

Xinshen Zhang, Zhen Ye, and Xu Zheng. Towards omnidirectional reasoning with 360-r1: A dataset, benchmark, and grpo-based method. *arXiv preprint arXiv:2505.14197*, 2025b. 3, 4

Xu Zheng, Chenfei Liao, Ziqiao Weng, Kaiyu Lei, Zihao Dongfang, Haocong He, Yuanhuiyi Lyu, Lutao Jiang, Lu Qi, Li Chen, Danda Pani Paudel, Kailun Yang, Linfeng Zhang, Luc Van Gool, and Xuming Hu. Panorama: The rise of omnidirectional vision in the embodied ai era. *arXiv preprint arXiv:2509.12989*, 2025. 2

Yikang Zhou, Tao Zhang, Dizhe Zhang, Shunping Ji, Xiangtai Li, and Lu Qi. Dense360: Dense understanding from omnidirectional panoramas. *arXiv preprint arXiv:2506.14471*, 2025. 2, 3, 4

Qirong Zhu, Ansheng Wang, Shinji Tanaka, Yasutoshi Makino, and Hiroyuki Shinoda. Vr-assisted guide dog training: A 360° panohaptic system for right-hand commands analysis. *arXiv preprint arXiv:2510.24057*, 2025. 2

A Data Pipeline Details

PanoArena uses separate construction pipelines for static images and dynamic videos. The image pipeline generates candidate QA pairs from complementary panoramic representations. The video pipeline combines geometry-based motion proposals, manually established object correspondences, and model-assisted event annotation. Model predictions serve as proposals for human verification.

A.1 Static Image Pipeline

Source and Filtering. We randomly sample one frame from each selected video in an Internet-collected ERP video pool to reduce repeated observations of the same scene. Images that cannot be decoded or do not have a $2:1$ aspect ratio are rejected before model inference. A locally deployed Qwen3.5-27B (Qwen Team, 2026) then checks two criteria: `is_erp`, whether the image exhibits the characteristics of an equirectangular panorama, and `is_informative`, whether its content is clear and informative. The informativeness criterion rejects nearly uniform, severely underexposed, or heavily obscured content. Images are retained when both checks pass. The structured verdict and a one-sentence reason are logged for inspection. The prompt is shown in Fig. 3.

Face Blurring. To protect privacy, we prepare face-blurred versions of the retained ERP images. For each image, we render six 640×640 cubemap faces and apply SAM3 (Carion et al., 2026) with the text prompt `human face` and a confidence threshold of 0.3. Each predicted face mask is filled using its convex hull in cubemap coordinates and projected back to ERP coordinates. The projected masks restrict Gaussian blurring to the corresponding face regions.

ERP–Cubemap Composite. To support recognition near the poles and across the ERP seam, we construct a labeled composite containing the full ERP image above two rows of cubemap views: left–front–right and top–back–bottom (Fig. 4). The ERP provides global orientation and scene topology, while the perspective faces reduce ERP distortion when inspecting local details. The center of the ERP image and the center of the front face both correspond to the camera's forward direction. This composite is the visual input to the annotation model.

QA Generation. Gemini-3.1-Pro-Preview (Google, 2026) receives the composite, task instructions, and manually annotated examples, and generates candidate QA pairs in one pass. The instructions cover the three static categories and six sub-tasks, with bearing-based and semantic-only target descriptions for Attribute and OCR. The questions must refer to the physical scene and avoid references to projection artifacts or composite panels; sub-tasks may be omitted when the image lacks suitable evidence. Direction, Attribute, and Layout use multiple-choice answers; Counting and OCR use open-ended answers. Each proposal includes a rationale and separate easy/hard difficulty labels for ERP and cubemap inputs, together with an indoor/outdoor scene label.

Human Verification. A dedicated web interface provides an interactive 360° viewer and the ERP–cubemap composite. First-round reviewers check target identity, question validity, options, answers, bearing-dependence labels, rationales, scene labels, and per-representation difficulty; they can edit or discard proposals and add questions about overlooked content. Questions whose visual evidence may be affected by face blurring, such as facial attributes or facing-direction judgments, are revised or removed when necessary. Second-round review reassesses the retained questions, with uncertain items referred for additional adjudication.

A.2 Dynamic Video Pipeline

Sources and Filtering. Motion reasoning uses raw ERP videos from an Internet-collected pool. Shot-boundary detection identifies continuous footage: the initial pool retains single-shot videos lasting 5–60 s, while an additional branch extracts 15–40 s segments from individual shots in longer videos. For dynamism screening, we sample the 30 fps source videos at 6 fps, resize the sampled frames to 320×160 grayscale pixels, and compute absolute differences between successive sampled frames. Each row is weighted by $\cos(\text{latitude})$, reducing the contribution of ERP's stretched polar regions. We retain candidates with high dynamism scores. Temporal perception uses paired ERP and perspective-rendered clips, as described below; its input preparation is separate from these motion-video filters.

```
You are an expert in ERP panoramic image quality assessment. Your task is to analyze
this image and determine if it is suitable for a high-quality equirectangular
panorama dataset.

Evaluate the image strictly based on the following 2 criteria. If either of these
criteria is not met, the final "keep" decision MUST be false:

1. is_erp: Does the image exhibit the standard characteristics of an Equirectangular
Projection (ERP) panorama?
2. is_informative: Is the image visually clear and informative? If the image is
predominantly solid black/white, extremely underexposed/dark making objects
indistinguishable, or heavily occluded by massive watermarks or subtitles, set this
to false.

Output ONLY a valid JSON object. Do not include any conversational filler,
explanations. Use the following structure:
{
    "is_erp": true/false,
    "is_informative": true/false,
    "keep": true/false,
    "reason": "A brief one-sentence reason for your decision"
}
```

Figure 3: **The quality filtering prompt in static image pipeline.**

Face Blurring. Faces in the retained ERP videos are blurred frame by frame using the same cubemap-based SAM3 procedure as in the image pipeline.

Target Relocalization. In the first stage, annotators identify the same physical object at two timestamps and use positive and negative clicks to obtain and refine SAM3 (Carion et al., 2026) masks in each frame. The annotator designates one observation as the reference and the other as the target, independently of their temporal order. The reference ERP frame, with the object outlined, identifies the object visually; the question asks for its camera-relative bearing at the designated target timestamp.

In the following geometric construction, subscripts i and j identify objects, while time dependence is written in parentheses. For object i at either annotated timestamp t, let $M_i(t)$ denote its set of mask pixels in an ERP frame of width W. For a pixel (u, v), u and v are the zero-based column and row coordinates measured from the top-left corner. The azimuth of column u is defined in radians as:

$$a(u) = \frac{2\pi u}{W} - \pi. \tag{2}$$

Zero azimuth points forward, and positive angles increase toward the camera's right. We estimate the object's representative azimuth $\theta_i(t)$, in radians, as the circular mean of its mask pixels' azimuths:

$$\theta_i(t) = \operatorname{atan2}\left(\sum_{(u,v)\in M_i(t)} \sin a(u), \sum_{(u,v)\in M_i(t)} \cos a(u) \right). \tag{3}$$

Here, $\operatorname{atan2}$ is the two-argument arctangent, which preserves the angle's quadrant. The target observation's azimuth is quantized into eight 45° sectors to determine the answer. We require at least 2 s between observations, at least 25° of bearing change, different source and target sectors, and a reference mask containing at least 300 pixels. Each question uses four alternatives drawn from either the cardinal or diagonal directions, according to the target sector.

Shared Geometry for Distance Trend and Motion Direction. We select clips with limited camera motion by measuring feature displacement between perspective views of sampled frames. SAM3 tracks movable object concepts, such as people and cars, on full ERP frames sampled at 2 fps within a nominal 10 s window in the clip.

The geometric construction combines an object's angular location, obtained from its segmentation mask, with its radial distance, estimated from depth. At each sampled time t, we compute the azimuth $\theta_i(t)$ of tracked object i from its mask $M_i(t)$ using Eq. 3. To recover its vertical location,



Figure 4: **An example of an ERP–cubemap composite.**

let H denote the ERP frame height and $N_i(t) = |M_i(t)|$ the number of pixels in its mask. The elevation $e(v)$ of pixel row v and the object's mean elevation $\phi_i(t)$, both in radians, are

$$e(v) = \frac{\pi}{2} - \frac{\pi v}{H}, \tag{4}$$

$$\phi_i(t) = \frac{1}{N_i(t)} \sum_{(u,v)\in M_i(t)} e(v). \tag{5}$$

Elevation is positive above the camera's horizontal plane.

Let $D(u, v, t)$ denote the radial depth predicted by DAP (Lin et al., 2026) at pixel (u, v) and time t. The median of valid depth values within $M_i(t)$ gives object i's estimated radial distance $r_i(t)$ from the camera. An object observation is skipped when too few mask pixels have valid depth. We then express the object's position as $\mathbf{p}_i(t) = (x_i(t), y_i(t), z_i(t))$ in a Cartesian coordinate system centered at the camera, with positive x, y, and z pointing right, up, and forward, respectively:

$$\mathbf{p}_i(t) = r_i(t)\big(\cos\phi_i(t)\sin\theta_i(t),\ \sin\phi_i(t),\ \cos\phi_i(t)\cos\theta_i(t)\big). \tag{6}$$

Applying this construction across sampled frames yields an estimated trajectory for each object, with task-specific smoothing applied before candidate generation. Qwen3.5-27B (Qwen Team, 2026) also generates short object descriptions from the video to provide referring expressions for the questions.

Distance Trend. The distance between object i and the camera is $d_{ic}(t) = r_i(t)$, where the subscript c identifies the camera. For two objects i and j observed at the same time, their pairwise distance is $d_{ij}(t) = \|\mathbf{p}_i(t) - \mathbf{p}_j(t)\|_2$, where $\|\cdot\|_2$ denotes the Euclidean norm. A five-sample median filter suppresses depth spikes. The automatic generator proposes four patterns: closer, farther, first closer then farther, and first farther then closer. Continuity checks reject excessive oscillation and abrupt jumps.

Motion Direction. Motion direction is the direction of local displacement in the camera's horizontal xz-plane. For object i, we median-smooth the horizontal position components and denote the resulting coordinates by $\tilde{x}_i(t)$ and $\tilde{z}_i(t)$. Among local intervals lasting 1.8–4.5 s, we select the interval $[t_0, t_1]$ with the largest net horizontal displacement for this object, where t_0 and t_1 are the interval's start and end times. The horizontal displacement components over this interval are

$$\begin{aligned}\Delta x_i(t_0, t_1) &= \tilde{x}_i(t_1) - \tilde{x}_i(t_0),\\ \Delta z_i(t_0, t_1) &= \tilde{z}_i(t_1) - \tilde{z}_i(t_0).\end{aligned} \tag{7}$$

For nonzero displacement, the motion angle $\psi_i(t_0, t_1)$, expressed in radians, is

$$\psi_i(t_0, t_1) = \operatorname{atan2}\big(\Delta x_i(t_0, t_1), \Delta z_i(t_0, t_1)\big). \tag{8}$$

We assign this angle to the nearest of eight sector centers using circular angular distance: 0° (front), 45° (front-right), 90° (right), 135° (rear-right), 180° (back), -135° (rear-left), -90° (left), and -45° (front-left). Each sector spans 45°, and the question refers to the interval midpoint, $(t_0 + t_1)/2$. For example, displacement along the camera's positive z-axis is labeled front regardless of the object's location in the panorama. Reviewers verify these direction proposals and use a not-moving label for stationary targets.

Temporal Perception. We generate event proposals from perspective-rendered counterparts of ERP clips and associate the resulting questions with the corresponding panoramic video. Frames are uniformly sampled at a nominal 2 fps, capped at 40 frames per clip and resized to a maximum side length of 640 pixels. In the model input, each frame is preceded by its timestamp, computed from its source-frame index and frame rate. Qwen3.5-27B (Qwen Team, 2026) returns event descriptions and start/end times, with instructions to describe people, objects, and actions and to avoid cinematography or viewpoint-motion descriptions. Post-processing constrains segment timestamps to the video duration, removes overlaps and degenerate intervals, and flags camera-related wording for inspection. The resulting segments need not cover the entire clip.

The retained event segments yield two types of proposals: Temporal Grounding asks when a described event occurs, and Event Recognition asks which event occurs in a specified interval. Event distractors preferentially come from other segments of the same video, supplemented by cross-video descriptions when needed. Temporal distractors use other event intervals or synthesized intervals with limited overlap with the correct span. Segment timings and distractor provenance are retained for human verification against the panoramic footage.

Human Annotation and Verification. Human processing comprises two stages, supported by task-specific tools that present panoramic playback, ERP–cubemap composites, perspective-rendered clips, or annotated frame pairs as appropriate. The first stage performs manual annotation for Target Relocalization: annotators establish object correspondences, designate reference and target observations, and refine their masks. For the other video tasks, the first stage checks and refines automatically generated QA pairs, including referring expressions, directional answers, distance trends,

Table 4: **Number of second-stage reviews per retained sample.** Counts refer only to samples in the final set. Each review is one assessment of a sample by one or two reviewers.

Sub-task	1 review	2 reviews	3 reviews	4 reviews	Samples
Target Relocalization	662	258	4	0	924
Distance Trend	251	178	7	0	436
Motion Direction	346	196	10	0	552
Event Recognition	197	100	18	0	315
Temporal Grounding	146	73	20	1	240
All	**1,602**	**805**	**59**	**1**	**2,467**

Table 5: **Second-stage agreement across all reviewed samples.** Each paired review contains one judgment from each of two reviewers. The *Record* column is the total number of paired review records; the next three columns partition that total into both approve, both reject, and disagree. *Agreement* is the sum of both-approve and both-reject counts divided by the total record number. Repeated paired reviews of one sample are counted separately.

Sub-task	Record	Approve	Reject	Disagree	Agreement
Target Relocalization	989	790	121	78	92.1%
Distance Trend	354	239	33	82	76.8%
Motion Direction	528	285	94	149	71.8%
Event Recognition	750	166	319	265	64.7%
Temporal Grounding	391	174	131	86	78.0%
All	**3,012**	**1,654**	**698**	**660**	**78.1%**

event descriptions, and temporal boundaries. Annotators can revise or discard proposals, add missing items, and assign scene labels and separate difficulty labels for ERP and cubemap inputs.

The second stage reviews the submitted annotations and QA pairs. For Target Relocalization, reviewers verify that the two observations depict the same object and that the masks accurately delineate it. For the other tasks, reviewers assess the validity of the questions and answers.

For all video tasks, the two stages form an iterative review–revision process. Items requiring correction are returned with feedback to the first-stage annotators, who can revise and resubmit them for second-stage review or discard unsuitable items; an item may be returned, revised, and reviewed more than once. For a subset of second-stage items, two independent reviewers assess each item.

The number of second-stage reviews can differ across retained samples. Table 4 summarizes the number of second-stage reviews recorded for each retained sample.

Table 5 reports all paired second-stage reviews in the review logs, including samples that were not retained in the final manifest. Agreement means that both reviewers approve or both reject the content, yielding 78.1% agreement across all 3,012 paired review records.

A.3 Final Formatting

The curated annotations are exported to modality-specific manifests sharing a common QA schema: sample identifier, task and sub-task, question, choices, answer, media path, scene, rationale, camera-bearing dependence label, and ERP/cubemap difficulty. Relocalization additionally supplies a reference-frame image. Open-ended questions have an empty choice list and a textual answer; multiple-choice questions use explicit option letters. Finally, option orders are deterministically rearranged within sub-task and option-count groups to reduce answer-position bias, with the corresponding answer letters updated consistently.

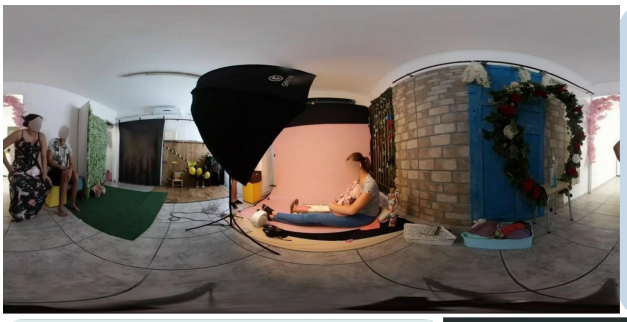
B Details of Evaluation Settings

When supported by a model, we compare three visual layouts: the full equirectangular panorama (ERP), perspective cubemap views (CMP), and a labeled ERP–cubemap collage (COL, as shown in Fig 4). The corresponding layout is used consistently for all frames of a video, with a fixed temporal sampling scheme and preserved timestamps.

Unless otherwise stated, evaluation is zero-shot with deterministic decoding (temperature 0 and one returned response per question). Models that expose a reasoning or thinking mode are evaluated in the reported mode, and only the final answer is used for scoring. Prompts, preprocessing, model checkpoints, and test manifests are frozen for each run; empty, invalid, or truncated responses remain in the evaluation denominator. Our token settings are followed the official recommendations of each model.

C PanoArena Sample Examples

Figs. 5–8 show examples of each sub-task of PanoArena. Blue and teal denote camera bearing-dependent and bearing-invariant questions, respectively, consistent with Fig. 1 and the other figures.



• Egocentric Direction

In which direction are the seated man and woman located relative to the camera?

A. To the physical right
B. Directly in front of the camera
C. Directly behind the camera
D. To the physical left

• Allocentric Direction

From the perspective of the man, in which direction is the steering wheel located?

A. To his physical right
B. Straight ahead
C. Behind him
D. To his physical left





• Attribute

What is the main feature of the wall located directly to the physical left of the camera?

A. A large mirrored wardrobe
B. Patterned wallpaper
C. A wall-mounted TV
D. A large window

• Attribute

What type of sports equipment is attached to the person's feet?

A. A snowboard
B. Skis
C. Ice skates
D. Snowshoes





• OCR

What is the four-letter word spelled out in large, multicolored letters on the ship?

AIDA

Figure 5: **PanoArena sample examples (part1).**

• Counting

How many go-karts are visible on the track in this scene?

3





• Layout

Which best describes the geometry of the paved path the vehicle is on?

A. It intersects another path to form a crossroad.
B. It forms a sharp 90-degree turn.
C. It forms a continuous straight line.
D. It splits into a Y-shape fork.

• Distance Trend

In the 0.0-6.5s of this video clip, how does the distance between the woman in a purple shirt climbing into a red kayak and the camera change?

A. first apart then closer
B. getting closer
C. unchanged
D. moving apart

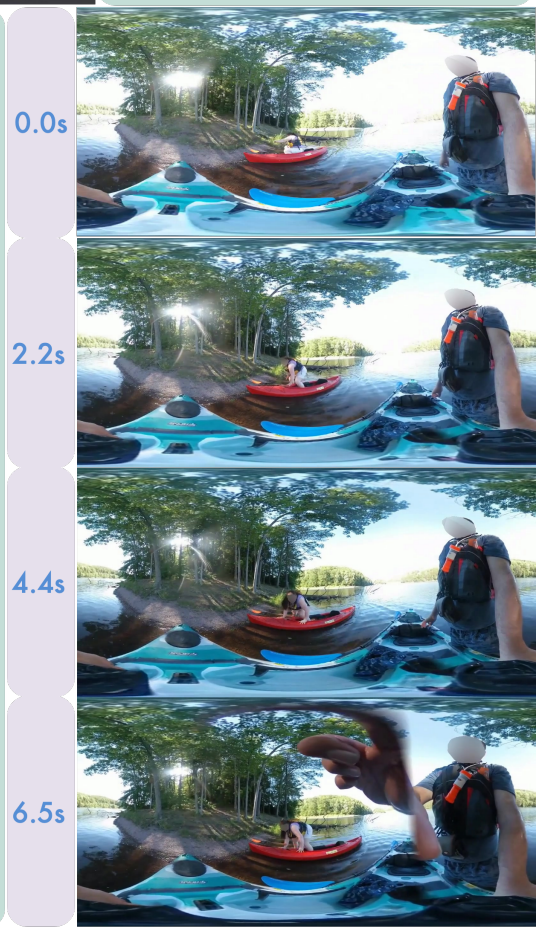


Figure 6: **PanoArena sample examples (part2).**

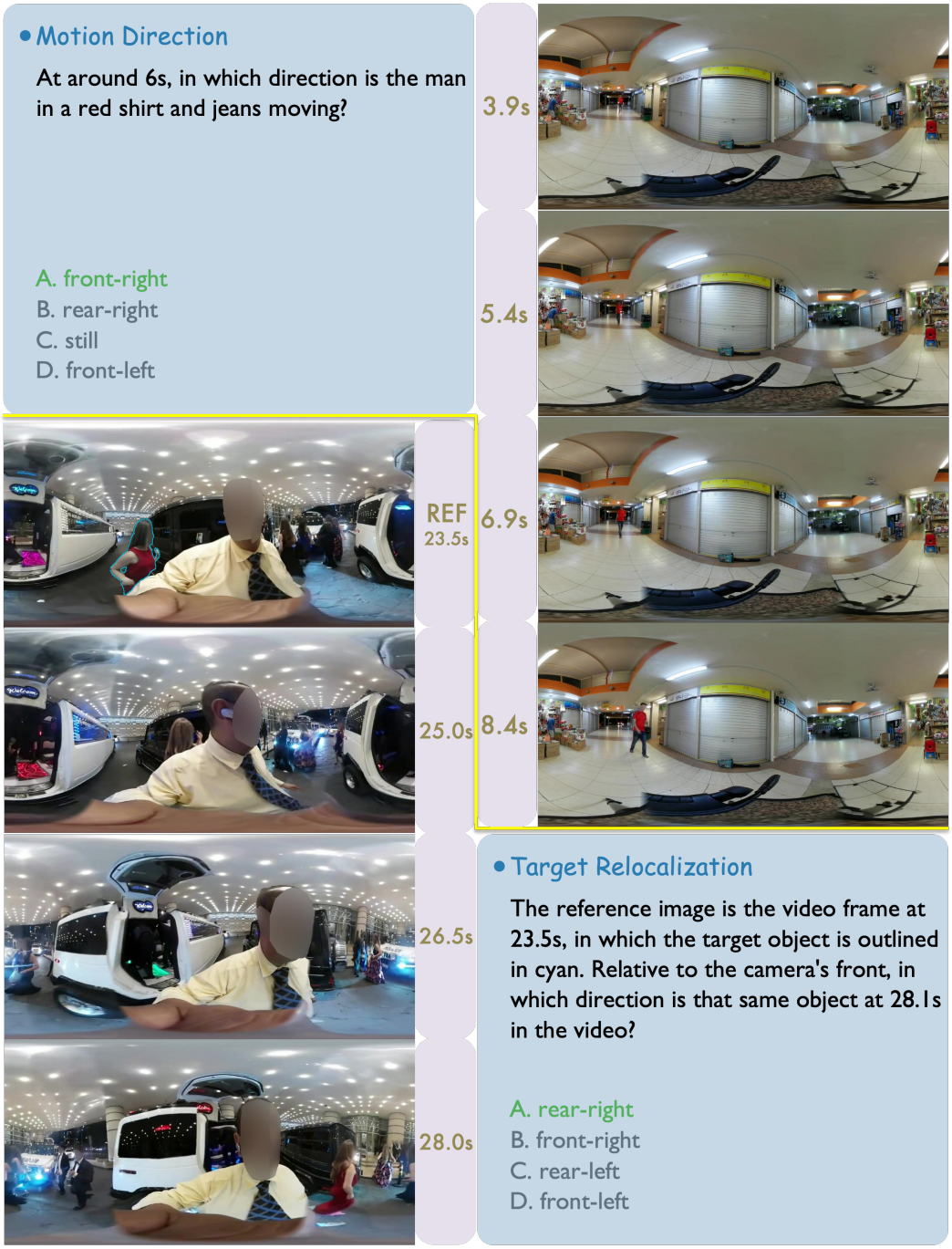


Figure 7: **PanoArena sample examples (part3).**

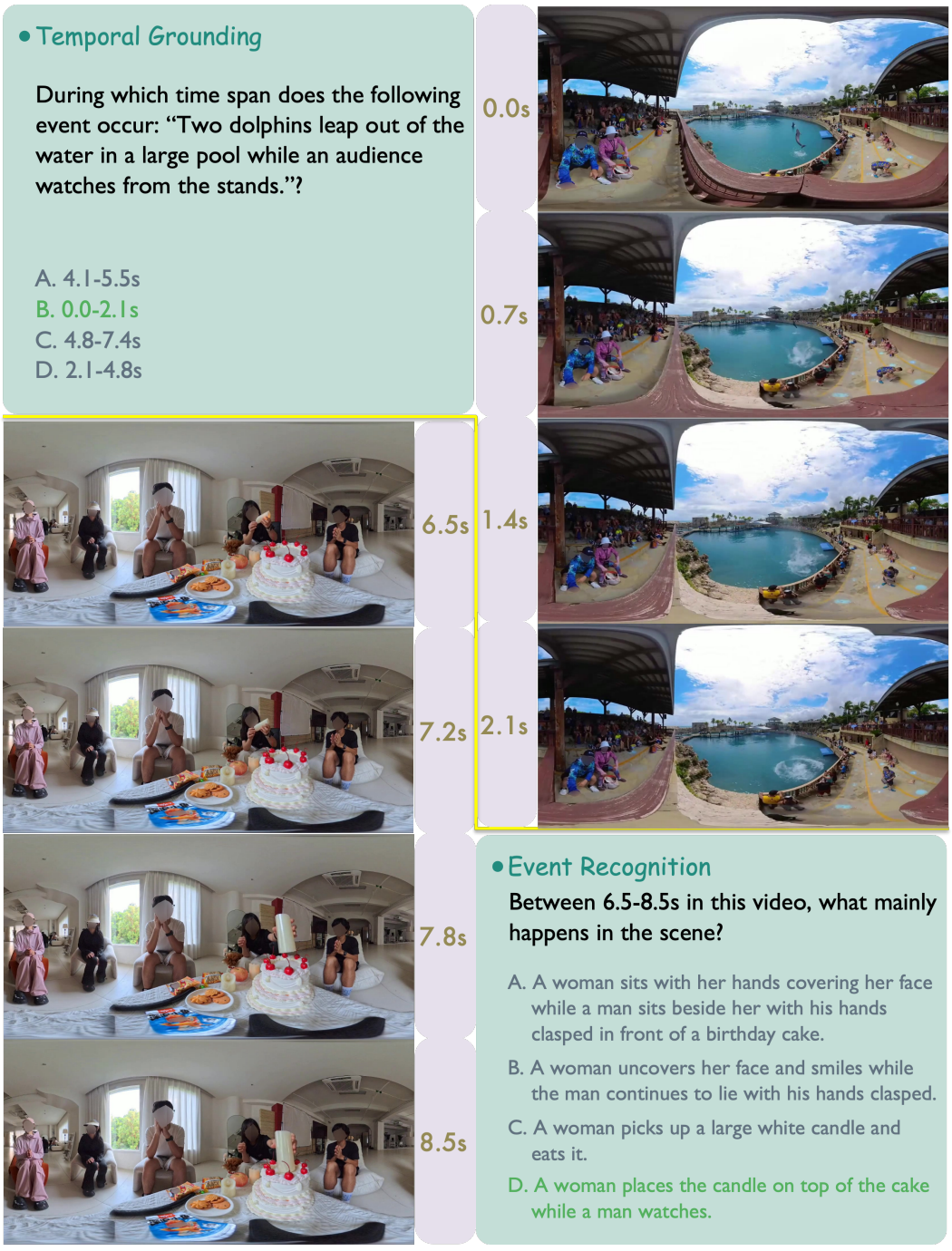


Figure 8: **PanoArena sample examples (part4).**