Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.DOI

TimePerceptBench: Evaluating Temporal Understanding in Multimodal Large Language Models with Remote Sensing Imagery

ERTAI E[1]

[1] Affiliation not provided

Corresponding author: Ertai E (e-mail: email@example.com).

ABSTRACT Temporal understanding represents a fundamental cognitive capability that remains limited in current multimodal large language models (MLLMs). While existing benchmarks primarily focus on static visual understanding or simple sequential patterns, they fail to adequately evaluate comprehensive temporal reasoning across extended sequences with objective ground truth. To address these limitations, we introduce TimePerceptBench, a comprehensive benchmark using timestamped remote sensing imagery with sparse temporal sampling (days to years). We explicitly distinguish two complementary capabilities: Temporal Order Understanding (TOU)—determining event sequences through 5 tasks, and Temporal Interval Understanding (TIU)—quantifying time intervals through 4 tasks, evaluated on 1,000+ multi-image groups. We evaluate 6 state-of-the-art MLLMs including GPT-4o mini and 5 open-source models. Unfine-tuned models show limited temporal reasoning: reordering achieves Positive-Negative Ratio (PNR) around 1.0–1.8, while TIU tasks perform near random baselines (interval category estimation: 20–23.5% vs 20%; extremum interval identification: 21–30% vs 25%). Fine-tuning substantially improves TOU (PNR up to 4.35), but TIU remains more challenging. This difficulty is not due to recognition deficits—models achieve 93.9% on UCM classification. Cross-domain validation shows TIU transfers better than TOU (+8.0% vs +3.9%), suggesting that TIU, despite its lower baseline, has greater room for improvement. TimePerceptBench establishes a systematic framework for temporal reasoning over discrete observations.

INDEX TERMS Temporal understanding, multimodal large language models, remote sensing, benchmark, temporal order understanding, temporal interval understanding.

I. INTRODUCTION

Temporal understanding—the ability to comprehend temporal relationships, sequential patterns, and duration estimation—represents a fundamental cognitive capability that enables humans to navigate the complexities of time-dependent events and dynamic environments. As multimodal large language models (MLLMs) achieve remarkable success in static visual understanding tasks such as image captioning, visual question answering, and object recognition Liu23 Yin24 Li24, their relatively weak temporal reasoning capabilities have gradually gained attention. This limitation becomes increasingly concerning as MLLMs are deployed in real-world applications requiring temporal perception.

While research efforts such as TimeBench Chu23 and TRAM Wang23 have attempted to evaluate symbolic and commonsense temporal understanding in text domains, the multimodal domain faces unique challenges that text-based benchmarks cannot address. Unlike textual temporal reasoning, which relies on explicit linguistic temporal markers, visual temporal understanding requires models to process and correlate information across multiple images or video frames, extracting temporal relationships from non-explicit linguistic cues such as visual changes, motion patterns, or environmental transformations Chen21 Dhingra22. While recent research such as TemporalVQA Imam25 has introduced basic temporal order understanding and time-lapse estimation tasks based on image pairs, these evaluations remain limited in scope and interval, typically addressing only short-term temporal relationships or simple sequential patterns.

Current multimodal benchmarks face three critical limitations: lack of objective temporal ground truth (temporal annotations in general-domain content heavily depend on subjective judgment), absence of systematic multi-image sequence evaluation (most work focuses on single images or simple image pairs), and failure to distinguish different temporal reasoning capabilities (conflating temporal order understanding with temporal interval quantification). Remote sensing imagery emerges as an ideal solution: satellite acquisition protocols provide objective timestamps that enable reliable temporal evaluation across multiple intervals; compared to typical video frames, remote sensing images present inherently more complex visual content with rich geographic and environmental information that demands sophisticated spatial-temporal reasoning; moreover, while continuous video datasets are abundant, discrete temporal remote sensing datasets remain relatively scarce, making our contribution particularly valuable for advancing temporal understanding research in challenging real-world scenarios.

To address these critical limitations and advance temporal understanding in multimodal large language models, we propose TimePerceptBench—a benchmark for evaluating temporal understanding capabilities in MLLMs using remote sensing image sequences. Unlike existing work that treats temporal reasoning as a monolithic capability, we explicitly distinguish two complementary yet fundamentally different capabilities:

- **Temporal Order Understanding (TOU):** Evaluates models' ability to comprehend event sequences through 5 tasks (Pairwise Order Verification, Sequential Order Verification, Image Sequence Reordering, Temporal Position Localization, Temporal Anomaly Detection), assessed on over 1,000 multi-image groups spanning various temporal intervals;
- **Temporal Interval Understanding (TIU):** Evaluates models' ability to quantify time intervals through 4 tasks (Extremum Interval Identification, Interval-based Pair Ranking, Pairwise Interval Comparison, Interval Range Estimation), covering multiple temporal categories from days to years.

This TOU/TIU separation framework enables us to systematically diagnose models' strengths and weaknesses across different dimensions of temporal reasoning, providing clear directions for targeted improvements.

We conduct comprehensive evaluation of 6 representative models, including GPT-4o mini Achiam23 and 5 state-of-the-art open-source models: InternVL3.5-8B-Instruct Wang25, Ovis2.5-9B Lu25, Qwen3-VL-8B-Instruct Yang25 Bai25, MiniCPM-V-4.5 Yu25, and MiMo-VL-7B-RL Yue25. As illustrated in fig. 1, the evaluation results reveal significant limitations in current temporal understanding capabilities:

Unfine-tuned models show limited temporal reasoning: On complex Image Sequence Reordering tasks, models achieve Positive-Negative Ratio (PNR) of only around 1.0–1.8 (near random guessing), while TIU tasks perform near random baselines (interval estimation: 20–23.5% vs 20% random; extremum identification: 21–30% vs 25% random). Fine-tuning substantially improves TOU capabilities (PNR up to 4.35), but TIU remains fundamentally harder (post-tuning accuracy of only 25–70%, compared to 70–95% for TOU).

parsed_images/0b6c3fc3-80f2-49bd-9180-9a3d89f8

FIGURE 1. The performance of each model in TimePerceptBench.

Notably, this temporal understanding difficulty is not due to visual recognition deficits—models achieve 93.9% accuracy on UCM land-use classification. Furthermore, cross-domain validation shows TIU transfers better than TOU (+8.0% vs +3.94%), suggesting that TIU, though harder to learn initially, gains more from remote sensing's temporal diversity, highlighting the unique value of remote sensing data for temporal understanding research.

Our benchmark makes several key contributions:

1) Comprehensive evaluation framework for MLLM temporal understanding using objective remote sensing timestamps, eliminating subjective annotation biases;
2) A systematic evaluation method that explicitly distinguishes TOU and TIU capabilities, through 9 complementary tasks (5 TOU + 4 TIU) covering multiple temporal intervals from days to years;

3) Revealing fundamental limitations in current temporal reasoning: TIU is more challenging than TOU, and this difficulty stems not from visual recognition deficits but reflects the inherent complexity of temporal quantification reasoning;
4) Deep insights into fine-tuning effects and cross-domain transfer: Fine-tuning significantly improves TOU but has limited impact on TIU, and TIU demonstrates better generalization in cross-domain scenarios, providing clear directions for future research.

II. RELATED WORKS

A. TEMPORAL REASONING BENCHMARKS FOR LARGE LANGUAGE MODELS

In recent years, the evaluation of temporal reasoning in language models has undergone significant development, with researchers developing increasingly sophisticated benchmarks to assess different aspects of temporal understanding. TimeQA Chen21 introduced time-sensitive question answering tasks focusing on temporal knowledge probing, demonstrating difficulties in time-sensitive information retrieval by models. TEMPLAMA Dhingra22 specifically targeted temporal knowledge by probing language models' understanding of temporal facts, revealing significant gaps in temporal knowledge representation. TEMPREASON Tan23 advanced the field by evaluating complex temporal relationship understanding, showing that even advanced models struggle with implicit temporal reasoning tasks.

Building upon these foundations, more comprehensive benchmarks have emerged. TRAM proposed the most extensive evaluation to date, containing 526.7k questions spanning ten different temporal reasoning tasks covering order, arithmetic, frequency, and duration aspects. Their systematic evaluation revealed that even the best-performing models, including GPT-4, fall significantly short of human-level performance, with GPT-4 Achiam23 achieving 84.4% accuracy while human performance exceeds 94%. TimeBench introduced a hierarchical framework categorizing temporal reasoning into symbolic, commonsense, and event temporal reasoning across 16 subtasks, demonstrating substantial performance gaps between state-of-the-art large language models and human capabilities. Their analysis revealed that models particularly struggle with computational, conversion, and comparison errors in temporal expressions, with multi-step reasoning remaining a significant challenge. Test of Time Fatemi24 specifically evaluates semantic understanding and arithmetic computation capabilities in temporal reasoning through synthetic datasets, avoiding issues with models leveraging prior knowledge, while PAT-Questions Meem24 focuses on present-anchored temporal question answering, addressing the dynamic challenges of answers changing over time and providing an automatically updated evaluation framework.

B. MULTIMODAL TEMPORAL UNDERSTANDING

Despite significant progress in text-based temporal reasoning, multimodal temporal understanding remains a significantly underexplored domain. Most existing multimodal benchmarks focus on static image understanding or simple video tasks that fail to adequately evaluate temporal reasoning capabilities across extended sequences Cai24. The transition from textual to visual temporal reasoning introduces unique challenges, as models must extract temporal relationships from visual changes rather than explicit linguistic cues, requiring fundamentally different cognitive processes.

Recent pioneering efforts have begun addressing this critical gap through specialized multimodal benchmarks. TOMATO Shangguan24 evaluates multimodal large language models' video understanding through comprehensive multi-frame analysis and frame order sensitivity assessment, focusing on sequential video frame processing and temporal consistency. Their evaluation revealed significant difficulties in temporal sequence understanding among current multimodal models, particularly in maintaining consistency across longer video sequences. TVBench Cores24 advances video-language evaluation by systematically addressing bias and static cue reliance in video-language tasks, advocating for more rigorous temporal evaluation protocols capable of distinguishing genuine temporal understanding from superficial pattern recognition.

Most directly relevant to our work, TemporalVQA introduced temporal order understanding and time-lapse estimation tasks using image pairs, representing the first systematic attempt to evaluate temporal reasoning in multimodal language models using static image sequences. Their comprehensive evaluation revealed striking limitations in current multimodal large language models' temporal reasoning capabilities, with advanced models like GPT-4V achieving only 49.1% accuracy on temporal order tasks and even lower performance on time-lapse estimation. This work demonstrated that the transition from text-based temporal reasoning to visual temporal reasoning represents a fundamental challenge for current multimodal architectures, highlighting the urgent need for more comprehensive evaluation frameworks capable of systematically assessing temporal understanding across diverse scenarios and extended temporal sequences.

Current temporal reasoning benchmarks exhibit several key limitations that constrain their effectiveness in comprehensively evaluating temporal understanding capabilities. Most existing evaluations primarily focus on short-term temporal relationships or simple sequential patterns, failing to capture the complexity of extended temporal reasoning required in real-world applications involving long-term planning and understanding. The reliance of many benchmarks on subjective temporal annotations or synthetic data limits evaluation objectivity and ecological validity, potentially introducing biases that affect assessment reliability and generalization to practical scenarios.

III. TIMEPERCEPTBENCH: TASKS AND DATASETS

TimePerceptBench provides a comprehensive evaluation framework for temporal understanding capabilities in multimodal large language models through two complementary tasks that capture fundamental aspects of temporal reasoning. Our benchmark design leverages remote sensing imagery to provide objective temporal ground truth while systematically assessing model performance across multiple temporal scales and complexity levels. These tasks are specifically designed to isolate different aspects of temporal cognition, enabling fine-grained analysis of model capabilities and limitations.

A. TASK OVERVIEW

Drawing inspiration from human temporal perception, we decompose temporal understanding into two core capabilities that reflect different cognitive processes involved in temporal reasoning. Temporal Order Understanding (TOU) evaluates the ability to comprehend chronological sequences and identify temporal relationships within multi-image sequences, representing the directional aspect of temporal cognition. Temporal Interval Understanding (TIU) evaluates the ability to comprehend time spans and identify temporal interval relationships between image pairs, representing the quantitative dimension of temporal understanding—analogous to how vectors possess both direction and magnitude. The multiple subtasks under these two tasks enable comprehensive evaluation while allowing for more systematic analysis of different temporal understanding mechanisms.

B. TEMPORAL ORDER UNDERSTANDING (TOU)

The Temporal Order Understanding task evaluates models' ability to comprehend chronological sequences and identify temporal relationships within multi-image sequences. This task addresses a fundamental aspect of temporal cognition—understanding how events unfold over time and their sequential relationships. Our TOU task is constructed using the Satellite Images Time Series Change Caption (SITSCC) dataset, which contains 1,000 groups of remote sensing image sequences. Each group captures the same geographical location across multiple time points, with each sequence containing an average of 5 or more images, covering various temporal scales from seasonal changes to multi-year urban development. We randomly partition the dataset into 800 groups for training and 200 groups for evaluation.

The TOU task encompasses five complementary subtasks at different granularities and difficulty levels, evaluating distinct aspects of temporal sequence understanding:

1) **Temporal Anomaly Localization (TAL)** presents models with a sequence of five images where four images from the same location form a coherent temporal progression, while one anomalous image either exhibits temporal inconsistency or originates from a different location, disrupting the evolutionary logic. Given a sequence $S = \{I_1, I_2, I_3, I_4, I_5\}$ where one image I_a is anomalous, the model can readily identify its temporal position through conspicuous visual patterns and inconsistencies in temporal evolution, outputting its temporal index. This task primarily evaluates the model's ability to recognize the positional identification of specific time points within a sequence.
2) **Image Sequence Reordering (ISR)** presents models with shuffled image sequences and requires them to predict the correct temporal ordering. Given an image sequence $S = \{I_1, I_2, \ldots, I_n\}$ with true temporal order O_{true}, the model receives a shuffled version S_{shuffled} and must predict an ordering O_{pred} that approximates the ground truth order O_{true}. This task evaluates the model's ability to understand temporal structure and sequential evolution patterns across time periods.
3) **Sequential Order Verification (SOV)** assesses whether models can distinguish between temporally consistent and inconsistent image sequences. Given an image sequence $S = \{I_1, I_2, \ldots, I_n\}$, the order may be temporally consistent (e.g., $I_1 \to I_2 \to \cdots \to I_n$) or shuffled (e.g., $I_3 \to I_1 \to \cdots \to I_n$). Models are presented with both correctly ordered sequences and shuffled sequences, requiring them to determine whether the sequence represents a reasonable temporal evolution process, outputting True or False. This binary classification task tests the model's sensitivity to temporal inconsistencies.
4) **Pairwise Order Verification (POV)** is a simplified version of SOV, focusing on local temporal relationships by presenting image pairs and requiring models to determine their temporal order. Given an image pair (I_i, I_j) where the true temporal order satisfies $t_i < t_j$, the model must judge whether the temporal sequence presented by these two images is reasonable, outputting True or False. This task evaluates the model's ability to identify direct temporal relationships and understand subtle temporal transitions between adjacent time points.
5) **Temporal Position Localization (TPL)** evaluates the model's ability to infer missing temporal positions within a structured sequence. Models are presented with four images: the first three images correspond to positions 1, 3, and 5 in a complete, evenly spaced temporal sequence, while the fourth image represents a missing time point. This task is a simplified version of ISR, where the model does not need to consider all positions for all images, but only needs to assess the matching degree between one image and two possible position orderings.

The TOU task provides comprehensive evaluation of temporal sequence understanding capabilities across multiple difficulty levels and reasoning granularities.

C. TEMPORAL INTERVAL UNDERSTANDING (TIU)

The Temporal Interval Understanding task evaluates the ability to comprehend time spans and identify temporal interval relationships between image pairs. This task addresses the quantitative aspect of temporal reasoning. Our TIU task utilizes the fMoW dataset, which contains over one million images from more than 200 countries with detailed metadata including precise timestamps. We select representative categories including airport_hangar, airport_terminal, factory_or_powerplant, and other infrastructure types that exhibit temporal evolution patterns. From each category, we extract image sequences from the same geographical locations, generating over 30,000 image pairs spanning different temporal scales, while filtering out image groups potentially affected by cloud cover or other interference based on their metadata. During the sample construction process, to avoid bias in model learning caused by uneven category distribution, we further filter and resample based on change detection results, ensuring that each TIU subtask has 800 groups for training and 200 groups for evaluation.

The TIU task encompasses four complementary subtasks at different granularities and difficulty levels, evaluating distinct aspects of temporal interval understanding:

1) **Interval Category Estimation (ICE)** evaluates the model's ability to estimate absolute temporal durations for image pairs. Given two images I_i and I_j of the same location taken at times t_i and t_j, the model must classify the time span $\Delta t = |t_j - t_i|$ into one of five categories: A. Days (1–30 days), B. 1–3 Months (31–90 days), C. 3–12 Months (91–365 days), D. 1–2 Years (366–730 days), or E. 2+ Years (731 days or more). This task evaluates the model's ability to perceive temporal magnitude from visual changes and map observed variations to discrete temporal categories.
2) **Pairwise Interval Comparison (PIC)** focuses on relative temporal interval judgment by presenting two image pairs from the same location at different times. Given four images I_a, I_b, I_c, I_d corresponding to times t_a, t_b, t_c, t_d, where (I_a, I_b) are from the same location and (I_c, I_d) are from another location, the model must determine whether $|t_b - t_a|$ is larger than $|t_d - t_c|$, outputting True or False. This task evaluates the model's ability to compare temporal intervals without requiring absolute quantification, mitigating the challenge that absolute time is difficult to judge in remote sensing imagery.
3) **Interval-based Pair Ranking (IPR)** extends PIC by requiring models to rank three image pairs by their temporal spans. Given three groups of image pairs A, B, and C, where each group shows the same location at different times, the model must rank these groups from shortest to longest time span. This task evaluates the model's ability to establish a global temporal ordering across multiple independent span observations, requiring simultaneous comparison and consistent ranking.
4) **Extremum Interval Identification (EII)** is a simplified version of IPR, assessing the model's ability to identify extreme temporal intervals among multiple options. Given four groups of image pairs—Group A through Group D—where each pair shows the same location at different times, the model only needs to identify which group has the longest time span without ranking all groups by their temporal spans.

The TIU task provides comprehensive evaluation of temporal interval understanding capabilities across absolute and relative temporal reasoning paradigms.

D. EVALUATION METRICS AND PROTOCOLS

Our evaluation framework employs task-specific metrics designed to capture different aspects of temporal understanding. For the reordering task under TOU, we utilize the Positive-Negative Ratio (PNR) as the primary evaluation metric, while for other binary judgment or classification tasks, we use accuracy as the evaluation metric, similar to the approach taken for TIU. The PNR metric can measure the stability and consistency of models in local temporal order judgment. Given an image sequence $S = \{I_1, I_2, \ldots, I_n\}$ with true temporal order O_{true} and model-predicted ordering O_{pred}, for any two images I_i and I_j in the same sequence, if they satisfy the same magnitude relationship in both the predicted ordering and true ordering, i.e., $(\text{rank}_{\text{pred}}(I_i) - \text{rank}_{\text{pred}}(I_j)) \cdot (\text{rank}_{\text{true}}(I_i) - \text{rank}_{\text{true}}(I_j)) > 0$, then this sample is recorded as a positive pair (Positive); if this product is negative, it is recorded as a negative pair (Negative).

The Positive-Negative Ratio is defined as follows:

$$\text{PNR} = \frac{N_{\text{positive}}}{N_{\text{negative}}} \tag{1}$$

where N_{positive} and N_{negative} represent the number of image pairs in the sequence that satisfy positive and negative order relationships, respectively.

The PNR metric exhibits higher robustness and interpretability compared to "full sequence accuracy." Unlike full sequence accuracy, which ignores potential local consistency in ordering, PNR provides more detailed measurement of model performance in order understanding by statistically analyzing the relative order relationships between all image pairs. This makes it particularly suitable for boundary cases in remote sensing imagery where visual ambiguity and unclear ordering may exist.

IV. EXPERIMENTS AND RESULTS

V. EXPERIMENTS AND RESULTS

We conducted comprehensive experiments on TimePercept-Bench to systematically evaluate the temporal understanding capabilities of current state-of-the-art multimodal large language models. Our evaluation encompasses six representative models selected for their exceptional visual processing capabilities: GPT-4o mini and five advanced open-source models, including InternVL3.5-8B-Instruct, Ovis2.5-9B, Qwen3-VL-8B-Instruct, MiniCPM-V-4.5, and MiMo-VL-7B-RL.

1) Experimental Setup

During data preprocessing, we uniformly resized all images to 256×256 resolution to ensure compatibility across different model architectures while preserving the critical visual information necessary for temporal reasoning. Each subtask of TOU and TIU consists of 800 training samples and 200 test samples. For fine-tuning experiments, we adopted LoRA (Low-Rank Adaptation) as a parameter-efficient fine-tuning strategy with carefully tuned hyperparameters: rank $r = 8$, scaling factor $\alpha = 16$, and dropout rate of 0.1. The training process employed a learning rate of 5×10^{-5} with a cosine annealing scheduler, batch size of 2, and gradient accumulation steps of 8. Models were fine-tuned for three epochs on the designated training set. Due to API limitations, GPT-4o mini was evaluated only in its baseline configuration without fine-tuning.

We evaluated models under two primary experimental conditions: (1) **Baseline:** original pretrained models without additional fine-tuning, and (2) **Fine-tuned:** models fine-tuned on our temporal reasoning tasks using a comprehensive temporal task training set composed of all subtask training sets, ensuring models do not exhibit particular bias toward any specific task and maintaining balanced temporal understanding.

For reordering tasks, we used the aforementioned PNR as the primary evaluation metric, while other selection or judgment tasks employed accuracy as the metric.

To validate the generalizability and transferability of learned temporal reasoning capabilities, we conducted additional cross-domain experiments on TemporalVQA, a non-remote-sensing benchmark (with image spans significantly different from our remote sensing datasets) that includes two tasks: image pair order judgment and image pair span prediction. This validation enables us to assess whether improvements observed on TimePerceptBench reflect genuine temporal understanding capabilities rather than dataset-specific overfitting.

2) Baseline Performance

Our evaluation reveals that unfine-tuned models demonstrate limited temporal reasoning capabilities on both TOU and TIU tasks, highlighting the challenges current multimodal large language models face in temporal understanding. Table 1 presents comprehensive baseline results for all six evaluated models across nine temporal reasoning subtasks (five TOU tasks and four TIU tasks).

For TOU tasks, baseline models exhibit highly unbalanced performance distributions. In the TAL task, most models demonstrate capabilities significantly above random baseline, with Ovis achieving 85.0% accuracy, and MiniCPM and MiMo-VL reaching 75.5%, indicating that models possess fundamental capabilities for localizing specific anomalous frame positions within temporal sequences. However, in the more challenging ISR task, baseline models achieve PNR scores mostly close to 1.0, approaching random ordering levels, with InternVL at only 1.0202, indicating models struggle to infer complete temporal orders from visual observations. Notably, GPT-4o mini achieves a PNR of 0.3093 on the ISR task, significantly below 1.0, which may reflect systematic misunderstanding of the ordering direction in task instructions. In other judgment-based TOU subtasks, model performance slightly exceeds random baselines: SOV task accuracy ranges from 53.0%–66.0% (random baseline 50%), POV task accuracy ranges from 45.5%–58.5% (random baseline 50%), and TPL task accuracy ranges from 53.5%–61.0% (random baseline 50%). These results indicate that while models can perform temporal judgments in simplified binary classification scenarios, their performance advantages remain relatively limited.

For TIU tasks, baseline performance reveals greater challenges in quantitative temporal reasoning. In the ICE task, model accuracy only slightly exceeds the 20% random baseline, indicating models struggle to map visual changes to discrete time intervals. The EII task similarly shows limited performance, with model accuracies approaching the 25% random baseline. Similar to ISR in TOU tasks, GPT-4o mini appears to misunderstand ordering direction in the IPR task's PNR scores in TIU, while other models similarly approach or fall slightly below random ordering. Model performance on the PIC task either slightly exceeds or falls slightly below random choice. These results indicate that without specific training, current multimodal large language models lack the ability to perform fine-grained temporal interval quantification from visual observations, with performance across all TIU subtasks clustering near random baselines.

3) Fine-tuned Model Performance

Fine-tuning on TimePerceptBench training data significantly enhances temporal reasoning capabilities, though improvement levels vary noticeably across different tasks. Table 2 presents fine-tuning results for five open-source models, demonstrating their learning capabilities on temporal reasoning tasks.

TABLE 1. Baseline Performance of Six Models

2*Model	TOU					TIU			
	TAL	ISR (PNR)	SOV	POV	TPL	ICE	PIC	IPR (PNR)	EII
Intern-VL	21.0%	1.0202	53.0%	45.5%	58.0%	20.0%	44.0%	0.8519	23.5%
Mimo-VL	75.5%	1.2497	57.5%	54.0%	61.0%	23.5%	55.5%	[illegible]	[illegible]
MiniCPM	75.5%	1.8409	60.5%	58.5%	58.5%	21.0%	44.5%	[illegible]	[illegible]
Ovis	85.0%	1.1299	53.0%	48.0%	53.5%	22.5%	56.5%	[illegible]	[illegible]
Qwen3-VL	66.5%	1.2989	66.0%	57.0%	56.0%	23.5%	56.5%	[illegible]	[illegible]
GPT-4o mini	78.5%	0.3093	57.5%	48.0%	61.0%	23.0%	46.0%	[illegible]	[illegible]

TABLE 2. Fine-tuned Model Performance

2*Model	TOU						
	TAL	ISR (PNR)	SOV	POV	TPL	ICE	PIC
Intern-VL	[illegible]	[illegible]	79.5%	65.5%	79.0%	26.5%	54.0
Mimo-VL	[illegible]	[illegible]	85.5%	76.0%	80.0%	31.5%	59.0
MiniCPM	[illegible]	[illegible]	84.5%	74.0%	80.0%	28.0%	53.0
Ovis	[illegible]	[illegible]	86.5%	77.0%	83.5%	30.0%	67.5
Qwen3-VL	[illegible]	[illegible]	90.5%	78.5%	83.0%	29.0%	70.0

For TOU tasks, fine-tuning produces significant improvements across all subtasks. Notably in the TAL task, the InternVL model's accuracy improves dramatically from 21% to 77.0%. While still exhibiting some gap compared to other models with accuracy above 90%, the 56 percentage point improvement indicates that the fine-tuned model has fully grasped the core mechanisms of this task. In the ISR task, fine-tuning brings the most remarkable improvements, with all models' PNR substantially increasing to above 2.75, with Ovis reaching 4.3476, nearly tripling from its baseline of 1.1299, indicating models successfully learned the ability to infer complete temporal orders from visual cues. Other TOU subtasks similarly exhibit consistent improvement patterns: SOV task accuracy improves to 79.5%–90.5% (baseline 53.0%–66.0%), POV task accuracy improves to 65.5%–78.5% (baseline 45.5%–58.5%), and TPL task accuracy improves to 79.0%–83.5% (baseline 53.5%–61.0%). These results indicate that with targeted training, models can establish systematic understanding of temporal order, achieving significant performance leaps across multiple granularity levels.

In contrast, TIU tasks show more modest improvements after fine-tuning, revealing the inherent difficulty of quantitative temporal reasoning. In the ICE task, while all models achieve performance improvements, the improvement margins remain relatively limited, with accuracy improving from baseline 20.0%–23.5% to 26.5%–31.5%, with the best model MiMo-VL reaching 31.5%, only an 8 percentage point improvement from its baseline of 23.5%. This limited improvement suggests that mapping visual changes to absolute temporal scales remains a fundamental challenge, difficult to completely overcome even with specialized training. On the PIC task, fine-tuning effects are relatively more pronounced, with accuracy improving to 53.0%–70.0% (baseline 44.0%–56.5%), with Qwen3-VL reaching 70.0%, a 13.5 percentage point improvement, indicating models possess stronger learning capabilities for relative temporal interval comparison. IPR task PNR scores improve to 1.29–2.16 (baseline 0.78–1.25); while improved, compared to the maximum PNR of 4.35 in the ISR task, ranking capability improvements in TIU tasks remain insufficient. In the EII task, accuracy improves from baseline 21.0%–30.0% to 27.5%–34.0%, with improvement margins of 4.5–7.5 percentage points, still significantly below ideal levels. These observations consistently indicate that TIU tasks, particularly those involving absolute temporal quantification (ICE and EII), pose more fundamental challenges to current model architectures.

Interestingly, we observe that baseline TOU and TIU performance are not perfectly correlated across different models. Models with stronger TOU baseline performance do not necessarily excel on TIU tasks, suggesting these capabilities may rely on partially independent cognitive mechanisms. Fine-tuning amplifies model-specific advantages while revealing systematic differences in temporal reasoning approaches across different architectures. Notably, GPT-4o mini, evaluated only in baseline configuration due to API limitations, exhibits performance levels comparable to open-source models, indicating that even advanced commercial models, when insufficiently large in scale and without task-specific fine-tuning, face fundamental challenges on certain temporal reasoning tasks.

4) Visual Recognition Capability Validation

To verify whether limited performance on temporal reasoning tasks stems from visual encoding deficiencies in models, we evaluated the visual recognition capabilities of five open-source models on the UCM (UC Merced Land Use) remote sensing image classification dataset. Table 3 presents the accuracy of all models on this standard visual classification task.

TABLE 3. Visual Recognition Capability Validation on UCM

Model	UCM Accuracy
Intern-VL	95.5%
Mimo-VL	95.5%
MiniCPM	93.0%
Ovis	94.0%
Qwen3-VL	91.5%

The UCM dataset contains 21 categories, and for each sample we select 5 categories including the correct category, with a random baseline of approximately 20%. Experimental results demonstrate that all evaluated models exhibit exceptional visual recognition capabilities, with accuracies all above 0.915, with InternVL and MiMo-VL achieving the highest accuracy of 0.955, Ovis reaching 0.94, MiniCPM reaching 0.93, and Qwen3-VL reaching 0.915. These results significantly exceed random guessing levels, demonstrating that all models possess robust remote sensing image feature extraction and category recognition capabilities.

This clearly excludes visual encoding deficiencies as the primary cause of limited temporal reasoning performance. The high accuracy of models on UCM (average 93.9%) contrasts sharply with their near-random baseline performance on TIU tasks (such as ICE and EII), indicating that temporal reasoning difficulties do not stem from models' inability to recognize or understand visual content in individual images, but rather from models' difficulty in establishing temporal relationships across multiple images and quantifying temporal intervals.

5) Cross-Domain Transferability Analysis

To evaluate whether temporal understanding capabilities learned on TimePerceptBench possess generalizability, we conducted cross-domain validation experiments on the TemporalVQA benchmark. TemporalVQA is a non-remote-sensing temporal reasoning benchmark containing image sequences of everyday life scenes, with time spans (with minimum units of seconds to minutes) significantly different from our remote sensing dataset (day to year levels). Table 4 presents the performance of five open-source models on TemporalVQA before and after fine-tuning, including evaluations across TOU (720 samples) and TIU (125 samples) dimensions.

TABLE 4. Cross-Domain Experiments on TemporalVQA, "Before" represents baseline models, "After" represents models fine-tuned on TimePerceptBench.

2*Model	TOU		TIU		Overall Accuracy	
	Before	After	Before	After	Before	After
Intern-VL	50.00%	58.61%	25.60%	50.40%	46.39%	57.40%
Mimo-VL	55.97%	61.11%	40.00%	53.60%	53.61%	60.00%
MiniCPM	67.50%	67.50%	56.80%	60.00%	65.90%	66.39%
Ovis	60.42%	65.56%	61.60%	64.80%	60.59%	65.44%
Qwen3-VL	60.83%	60.83%	63.20%	58.40%	60.47%	60.47%

Experimental results reveal interesting patterns in temporal reasoning capability transfer. Overall, models fine-tuned on TimePerceptBench demonstrate certain degrees of performance improvement on TemporalVQA, demonstrating that learned temporal understanding capabilities possess certain cross-domain generalizability. However, a more significant finding is that TIU capability transfer effects are notably superior to TOU capability transfer. TIU tasks improve by approximately 8.0 percentage points on average after fine-tuning (from baseline 25.6%–63.2% to 50.4%–64.8%), while TOU tasks improve by only approximately 3.9 percentage points (from baseline 50.0%–67.5% to 58.6%–67.5%). This difference is also confirmed at the individual model level.

This asymmetry in transfer patterns reveals essential differences between different dimensions of temporal reasoning. The stronger cross-domain transfer of TIU capabilities suggests that cognitive mechanisms underlying temporal interval understanding may be more universal and abstract, not strongly constrained by specific image domains or temporal scales. Knowledge learned by models about relationships between degrees of change and temporal spans can effectively transfer across different visual domains and temporal scales. In contrast, limited TIU transfer may reflect stronger dependence of temporal order understanding on domain-specific patterns—specific temporal patterns learned in remote sensing images, such as seasonal changes and urban development, are difficult to directly apply to different temporal logic in everyday scenes. Additionally, Task 1 (temporal order understanding) improves from baseline 46.39%–65.92% to 57.4%–66.39% after fine-tuning, with improvement margins between 10–20 percentage points, exhibiting transfer patterns similar to main TOU tasks. Notably, different models exhibit different characteristics in cross-domain transfer: InternVL achieves the largest transfer gains on TIU, while Qwen3-VL's TIU performance slightly decreases after fine-tuning (from 63.2% to 58.4%), potentially reflecting overfitting or model architecture preferences for specific temporal scales.

These cross-domain experimental results have important theoretical and practical implications. They demonstrate that temporal understanding capabilities learned on TimePerceptBench are not simple dataset overfitting, but possess certain degrees of generalization capability. Second, the stronger transferability of TIU compared to TOU, combined with its lower absolute performance on TimePerceptBench, suggests that TIU tasks, while more challenging, may possess greater room for improvement and stronger universal value.

VI. CONCLUSION

We present TimePerceptBench, a comprehensive benchmark for evaluating temporal understanding capabilities in multimodal large language models. By leveraging remote sensing imagery with objective timestamps, we systematically decompose temporal understanding into two complementary core capabilities: Temporal Order Understanding (TOU) and Temporal Interval Understanding (TIU), designing nine fine-grained subtasks to evaluate model performance across different granularities and difficulty levels.

We conducted comprehensive evaluation of six advanced multimodal large language models, with experimental results revealing significant limitations in current models' temporal reasoning capabilities. Unfine-tuned models perform near random baselines on temporal sequence reordering and temporal interval quantification tasks, indicating that temporal understanding capabilities are not inherent to these models. Through 93.9% high accuracy on the UCM dataset, we clearly exclude visual encoding deficiencies as the primary cause of limited performance, confirming the inherent challenges of temporal reasoning itself.

Fine-tuning experiments demonstrate that models can significantly improve temporal reasoning capabilities after targeted training, but TOU and TIU tasks exhibit different learning characteristics. TOU tasks show more pronounced improvements, with reordering tasks achieving PNR up to 4.35, while TIU tasks, especially absolute temporal quantification, remain highly challenging, indicating fundamental difficulties in current model architectures for quantitative temporal reasoning. Cross-domain validation experiments further reveal interesting transfer patterns: TIU capabilities demonstrate stronger cross-domain transferability compared to TOU (+8.0% vs +3.9%), suggesting that despite TIU tasks being more challenging, the cognitive mechanisms they rely upon may be more universal with greater potential for improvement.

TimePerceptBench provides a systematic evaluation framework and in-depth analytical insights for temporal understanding research. We hope this benchmark will advance further development of temporal reasoning capabilities in multimodal large language models and facilitate the construction of more intelligent and comprehensive visual understanding systems.