Orderbench: A Unified Benchmark for Temporal and Causal Reasoning Across Multimodal, World-Model, and Embodied AI Systems

Anonymous ACL submission

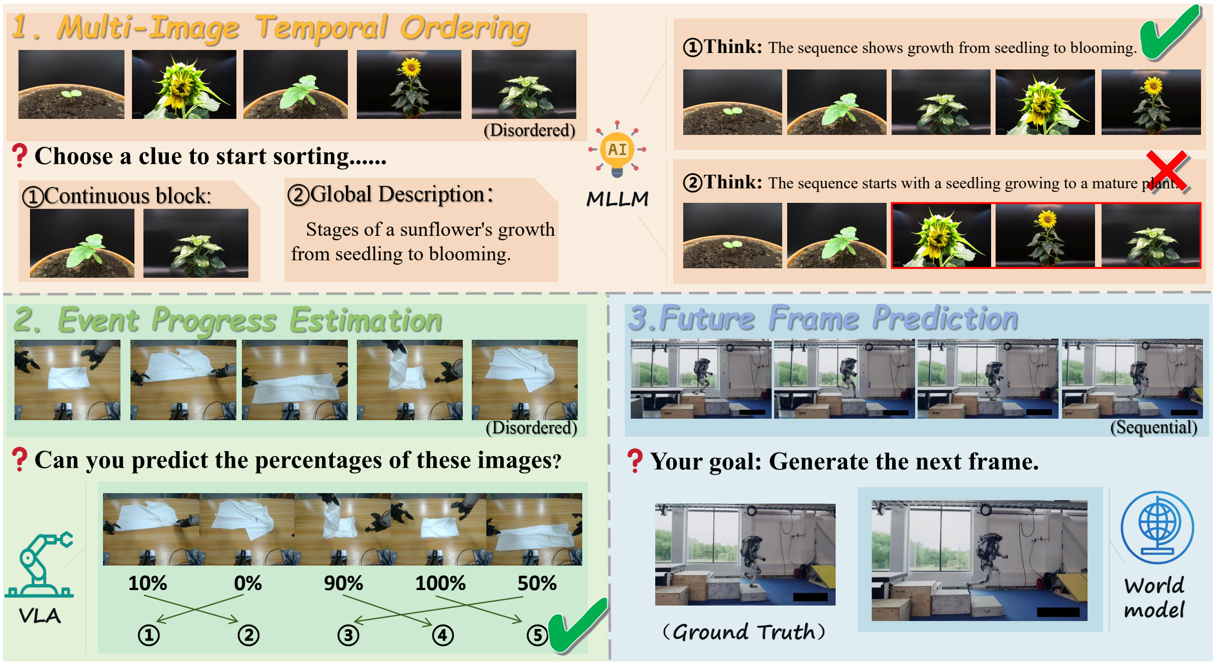

Figure 1: **Overview of the Orderbench benchmark.** Each sequence depicts coherent **real-world** events, and models are required to *reconstruct the correct order* or to *predict future frames* beyond the observed sequence. This setting enables a unified evaluation across three model families: ***i.*** **Multimodal Large Language Models**, which interpret visual order and semantic coherence through image–text reasoning; ***ii.*** **World Models**, which predict plausible future frames or dynamic transitions within continuous environments; ***iii.*** **Vision Language Action models**, which infer action progression and goal-oriented reasoning for embodied agents. Together, these paradigms make **Orderbench** a comprehensive benchmark for assessing perception, prediction, and embodied temporal understanding.

Abstract

Intelligent systems that operate in the real world, whether generating future predictions, executing embodied instructions, or interpreting complex visual scenes, share a common prerequisite: the ability to reason about *when* and *why* events occur. Despite this shared dependency, multimodal large language models (MLLMs), world models, and vision-language-action systems (VLAs) have historically been evaluated in isolation, with no unified framework capable of exposing their common temporal and causal reasoning limitations. We introduce **Orderbench** which identifies temporal ordering as the natural intersection across domains, enabling the first comparative evaluation of their temporal and causal reasoning capability through realistic event reconstruction from shuffled video frames. **Orderbench** adapts this core challenge to each model family's strengths: frame ranking for MLLMs, task progress estimation for VLAs, and future frame prediction for world models, creating an organically unified evaluation framework spanning 4,000 samples across daily life and robotics domains. Extensive experiments reveal that even the most advanced models struggle significantly, with best performance under 40% accuracy and 10-20% performance gaps between daily and robotic scenarios, exposing a critical disconnect in temporal cognition. Building on this observation, we further explore factors that elicit temporal and causal reasoning in current models. We believe this work will provide guidance for research on causally-aware world models and embodied AI systems.

1 Introduction

> *"Order is heaven's first law."*
>
> — Alexander Pope

The pursuit of intelligent systems capable of operating in the real world, from embodied robotic agents (Vision-Language-Action models, or VLAs) to sophisticated predictive world models, hinges on a fundamental capability: *reasoning about temporal dynamics and causality* (Gladyshev et al., 2025; Yu, 2025; Ma et al., 2024; Bi et al., 2025; Liu et al., 2025). World models aim to forecast future states from current conditions, while VLAs must select actions to achieve a goal. For a VLA to plan effectively, it must implicitly leverage a world model, understanding that action A must precede B to cause outcome C. Both lines of research, therefore, are deeply reliant on a fine-grained, robust understanding of temporal order and causal relationships in videos.

Despite this need, existing benchmarks fail to adequately assess temporal reasoning at the granularity required for real-world operation. Many video understanding evaluations can be solved through static frame-level perception alone, or remain limited to high-level event sequences (Sharma et al., 2022; Daniels and Metaxas, 2022; Lin et al., 2022; Girdhar and Ramanan, 2019; Yi et al., 2019). Prior work has further shown that several benchmarks exhibit temporal bias, where questions can be correctly answered from single or shuffled frames (Shangguan et al., 2024). They fail to probe the fine-grained temporal logic essential for complex planning and genuine physical understanding.

To rigorously assess this *latent* capability, we identify "video temporal ordering" as **the fundamental intersection** across MLLMs, VLAs, and world models—a shared challenge that naturally gives rise to a unified evaluation framework. We present models with shuffled video frames and require them to reconstruct the single, coherent event sequence. This task is deceptively simple for humans, who leverage a lifetime of world knowledge to instantly identify the correct flow of events. For current multimodal systems, however, it is exceptionally challenging (Yang et al., 2024; Hu et al., 2021; Li et al., 2025). Success demands more than visual recognition; it requires models to infer the plausible temporal and causal relationships between discrete steps (Wang et al., 2025b), bridging perception with rich, implicit world knowledge.

We operationalize this protocol in our new large-scale benchmark, **Orderbench**. The dataset consists of 4,000 meticulously curated samples, intentionally designed to evaluate a diverse range of reasoning skills. The benchmark encompasses two primary domains: Daily Life (covering sports, occupational skills, science, and everyday activities) and Robot (covering basic object manipulation, household tasks, precision kitchen skills and sports), which are further categorized into 74 distinct sub-categories.

Crucially, temporal ordering manifests differently across model families, reflecting their distinct architectural objectives and capabilities (as shown in Fig. 1). MLLMs excel at semantic-visual reasoning, VLAs operate in goal-conditioned embodied contexts, and world models focus on predictive dynamics. **Orderbench** naturally adapts the core ordering challenge to these specializations: *i*.MLLMs are evaluated via direct frame ranking, assessing their perception-driven understanding of visual order and semantic coherence. *ii*.VLAs are assessed through task progress estimation, a goal-aware temporal query that measures embodied, goal-conditioned reasoning. *iii*.World models are tested via future frame prediction, directly probing their ability to extrapolate temporal consistency and causal transitions. These task-specific instantiations of the same underlying challenge enable, for the first time, comparison of temporal reasoning across previously disparate model families.

Our comprehensive evaluation reveals a consistent and sobering picture. Even the strongest proprietary models fail to exceed 40% strict sequence accuracy. Performance degrades by 10–20% when shifting from daily-life to robotic scenarios, exposing a critical gap in cross-domain temporal generalization. Strikingly, many models can *accurately describe a pre-ordered event sequence yet fail dramatically at reconstructing* that same sequence from shuffled frames, revealing a fundamental disconnect between static scene understanding and dynamic temporal reasoning. Analysis of training-free enhancement strategies further shows that structured contextual information and explicit relational modeling are more effective than complex multi-step reasoning chains when model capacity is limited. Our contributions are as follows:

- We identify temporal ordering as the fundamental capability underlying MLLMs, VLAs, and world models, and introduce **Orderbench**

— the first benchmark enabling comparative evaluation across these previously disparate model families through task-specific adaptations of frame sequencing.

- We curate a large-scale dataset spanning daily life and robotics domains, with 4,000 carefully annotated samples across 74 subcategories, designed to probe fine-grained temporal and causal reasoning.

- Comprehensive evaluation reveals significant performance gaps, with best models achieving under 40% accuracy and exhibiting 10-20% worse performance on robotic versus daily scenarios, exposing critical limitations in cross-domain temporal reasoning.

- We evaluate a suite of training-free enhancement strategies and provide actionable insights into what contextual and relational signals most effectively elicit temporal reasoning in current models.

We hope **Orderbench** and our analysis will inspire future work in developing more sophisticated and causally-aware world models and embodied AI systems.

2 Related Work

Multi-image Understanding. This capability refers to the ability of model to integrate, compare, and reason across multiple related images, representing a crucial dimension for evaluating MLLMs. Traditional benchmarks such as MME (Zhang et al., 2021), MMBench (Liu et al., 2024), and MMMU (Yue et al., 2024) primarily focus on single-image question answering or description tasks. Even larger-scale benchmarks like MMIU (Patel et al., 2021) offer limited coverage of cross-image reasoning. Recent benchmarks have begun to address multi-image inputs: Q-Bench (Zhang et al., 2024) evaluates perceptual comparison, Memontos (Wang et al., 2024b) , STRIPCIPHER (Wang et al., 2025a) and TempVS (Song et al., 2025) extends to temporal sequences, while DEMON (Li et al., 2023) and MANTIS-Eval (Jiang et al., 2024) and MIR (Du et al., 2025) focus on multi-image instruction following. M4Bench (Ye et al.), MMRB (Cheng et al., 2025), BLINK (Fu et al., 2024), and MuirBench (Wang et al., 2024a)expand cross-domain reasoning and object consistency evaluation, but remain limited to static understanding without event-level validation. **Orderbench**fills this gap by reconstructing event sequences from shuffled frames for direct cross-image reasoning assessment.

Visual Temporal Reasoning. Visual temporal reasoning is a critical capability for models to comprehend event sequences and causal relationships, serving as an essential bridge from static perception to dynamic understanding. Existing approaches predominantly rely on video-based benchmarks such as VITATECS (Li et al., 2024b), MVBench (Li et al., 2024a), SIMS-VSI (Brown et al., 2025), VideoThinkBench (Tong et al., 2025), and ReXTime (Liao et al., 2024), which assess temporal consistency and causal reasoning. TOMATO (Shangguan et al., 2024) shows these benchmarks suffer from inherent temporal biases. To mitigate motion dependency, MMRB (Cheng et al., 2025) and TemporalVQA (Fazli Imam et al., 2025) adopt image-based evaluation without motion cues, showing MLLMs still struggle with temporal and causal reasoning. In contrast, **Orderbench**reconstructs temporal logic from unordered frames, mitigating video bias and offering a more rigorous evaluation.

Towards Unified Evaluation. The fragmented evaluation of MLLMs, VLAs, and world models under family-specific protocols has motivated recent efforts toward cross-paradigm assessment. UniSim-Bench (Ghazanfari et al., 2024) unifies 7 multimodal perceptual similarity tasks across 25 datasets, revealing that task-specialized models fail to generalize to unseen but related tasks. MultiNet v1.0 (Guruprasad et al., 2025) evaluates both VLMs and VLAs across six capability regimes, including visual grounding, spatial reasoning, tool use, physical commonsense, multi-agent coordination, and continuous control, finding that no current model demonstrates consistent cross-domain generality. On the VLA front, vla-eval (Choi et al., 2026) provides a unified evaluation harness decoupling model inference from 13 simulation benchmarks, enabling reproducible cross-model comparison with a standardized protocol. StarVLA (Community, 2026) further bridges VLM-based and world-model-based paradigms by accommodating both backbone types within a single modular architecture under shared benchmarks. Meanwhile, UNIVERSE (Hendriksen et al., 2025) adapts VLMs for fine-grained, temporally grounded eval-

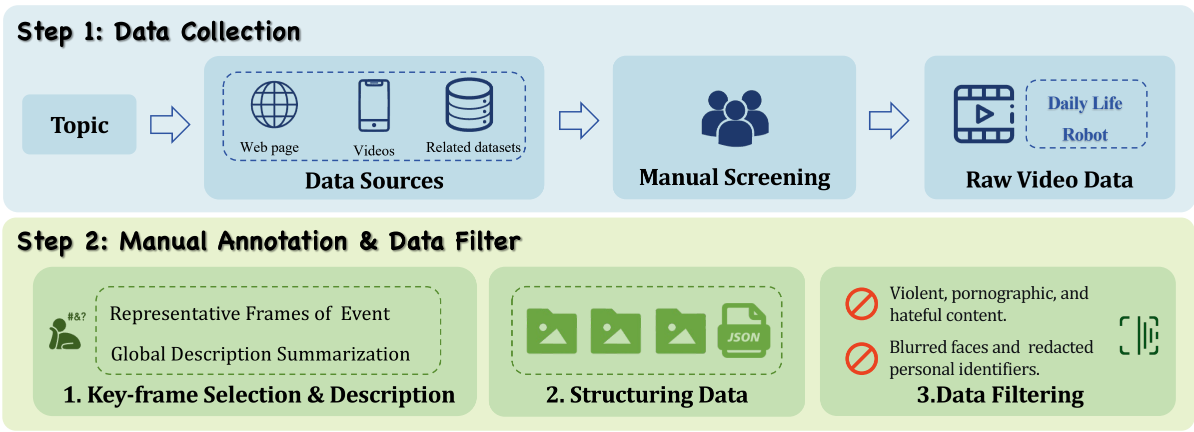

Figure 2: **Data composition and filtering pipeline.** We first collect raw videos from online and robotic sources, and then conduct human annotation and multi-stage quality filtering to ensure clear temporal progression and balanced domain diversity.

uation of world model rollouts, addressing a gap left by conventional metrics that cannot assess action alignment or semantic consistency over time. **Orderbench** fills this gap by identifying temporal ordering as the shared challenge, enabling cross-family comparison on a unified data foundation.

3 Orderbench

Building on the insight that temporal ordering represents the natural intersection across MLLMs, VLAs, and world models, **Orderbench** establishes a comprehensive temporal reasoning evaluation framework. The benchmark comprises 4,000 five-frame sequence samples spanning two complementary domains: daily life and robotic operations. Evaluation is conducted through three task paradigms tailored to different model families: multi-image temporal ordering, event progress estimation, and future frame prediction (detailed in Sec. 4). Fig. 2 illustrates the construction pipeline.

3.1 Data Collection

We construct a balanced dataset of 4,000 samples, evenly split between daily life and robotic domains. Each sample consists of a five-frame sequence, a length that provides sufficient temporal context for multi-step causal reasoning while maintaining a challenging combinatorial space of 120 possible orderings. An ablation over frame counts from 3 to 7 confirms that five frames yields the highest information entropy among configurations where model performance remains meaningfully above chance, as detailed in Appendix.

Daily Life Domain. We systematically collected more than 2,000 high-quality samples covering typical daily scenarios such as sports, cooking, and handicrafts. These scenarios require reasoning about object state changes, action sequences, and causal relationships grounded in commonsense world knowledge, providing a foundation for evaluating general temporal understanding.

Robot Domain. This domain is primarily sourced from curated video platforms and multiple publicly available real-world and simulated robotic operation datasets (e.g., Open X-Embodiment (O'Neill et al., 2024), MimicGen (Mandlekar et al., 2023), AgiBot (Bu et al., 2025); full list in appendix). These datasets cover scenarios including basic object manipulation, household tasks, fine-grained kitchen skills, and sports. This domain emphasizes embodied reasoning and goal-conditioned temporal understanding, testing capabilities essential for VLAs.

3.2 Human Annotation

To curate the **Orderbench** benchmark from our collected raw video data, we designed a human annotation pipeline to generate two critical components for each sample: a sequence of five representative *key frames* and a corresponding *global description* of the event.

We recruited 17 annotators (10 male, 7 female), all of whom are college students with a research background in computer vision, to perform this task. The annotation process followed a structured protocol: *i*.**Contextual Understanding**: Annotators first watched the entire original video to grasp the event context. *ii*.**Key Frame Selection**: They selected five representative frames that best capture the key nodes and salient moments of the event pro-

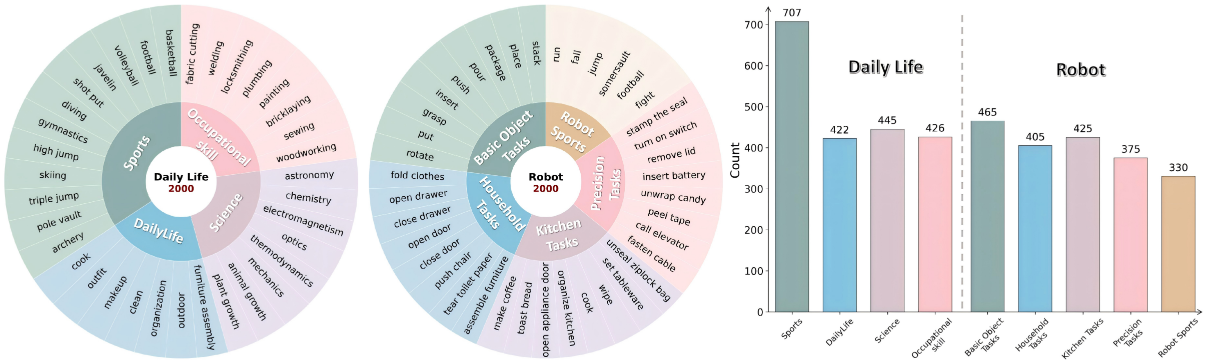

Figure 3: **Data distribution of Orderbench**. The benchmark consists of an equal proportion of *daily life* and *Robot* videos. The left panel illustrates the topic coverage for each domain, while the right panel presents the quantity for each category.

gression. *iii*.**Global Description Writing**: Based only on the visual content of these five selected frames, annotators composed a global description.

This global description was created under strict constraints. The text must accurately summarize the event, provide semantic context for temporal reasoning, and avoid any explicit ordering cues that would create ambiguity. The goal is to help a model understand the event's context without revealing the answer, isolating core reasoning capability.

3.3 Quality Control

To ensure the quality of the dataset, we applied a rigorous quality control process for all samples. Each sorting task was reviewed by the author team to verify that it met the following core criteria: *i*.Image content must be clearly identifiable, text descriptions must be accurate and free of any violent, pornographic, or other inappropriate content; *ii*.Sequences must not exhibit logical inconsistencies or temporal ambiguities. Any sample that did not meet these quality requirements was immediately discarded to maintain the overall quality.

We assessed annotation consistency on a randomly sampled subset. Four independent annotators re-sorted each sequence, achieving a mean inter-annotator Kendall's τ of 0.79 and a mean annotator-vs-ground-truth τ of 0.87, confirming that the established orderings are unambiguous and reproducible. Details are provided in Appendix.

3.4 Dataset Statistics

The **Orderbench** contains 4,000 multi-image sequence samples in total, divided into two domains: Daily Life (2,000 samples) and Robot (2,000 samples), as illustrated in Fig. 3. Each sample includes five key images and a text description. The metadata, including sample ID, category labels, subcategories, descriptive text, and image paths, is stored in a unified JSON format.

Daily Life Domain. For daily life samples, a four-level classification system was established: *i*. Sports: Various types of sports training and competitive events. *ii*. Daily Life: Routine interpersonal interactions and everyday activities. *iii*. Science: Demonstrations of physical principles and natural phenomena. *iv*. Professional operational skills and technical procedures. This classification was further refined into 35 subcategories to ensure diversity and representativeness.

Robot Domain. This domain focuses on embodied visual reasoning and agent–environment interaction, integrating single-view robot operation records, multi-view observations, and simulation environment data. A five-level classification system was adopted: *i*. Basic Tasks: Fundamental perception and grasping operations. *ii*. Household Tasks: Routine service tasks in domestic environments. *iii*. Kitchen Tasks: Tasks performed in kitchen scenarios. *iv*. Precision Operations: High-precision assembly and manipulation tasks. *v*. Robot Sports: Robot-specific dynamic and coordination tasks. The classification was further divided into 39 subcategories.

4 Evaluation Protocol

As temporal ordering differs across model families, we adapt this shared challenge into three task-specific protocols aligned with each architecture's capabilities: multi-image temporal ordering for MLLMs, event progress estimation for VLAs, and future frame prediction for world models.

4.1 Multi-Image Temporal Ordering for MLLMs

This task is designed to systematically evaluate MLLMs' capability to understand temporal event progression and recover the correct chronological sequence from visual observations. We provide the model with five images depicting different developmental stages of the same event, which are shuffled according to a predetermined unified scheme to ensure identical conditions for all models. The model's objective is to infer the plausible temporal order based on visual cues such as object state transitions, action continuity, and scene layout evolution. The core goal is to test the model's ability to construct a coherent visual narrative rather than simply performing isolated recognition of individual images. This direct ranking format leverages MLLMs' strengths in semantic-visual reasoning and cross-image comparison. To systematically analyze the role of prior knowledge in this temporal reasoning process, we designed three progressive evaluation settings.

No-Context Condition provides only the image sequence, requiring the model to reason based solely on visual information, thereby assessing its pure visual temporal understanding.

Global Description Condition provides a semantic description of the event in addition to the images, examining how the model integrates visual and semantic information.

Partial Ordered Block Condition provides image sub-blocks with known internal ordering, testing the model's ability to utilize local temporal priors to infer the complete sequence.

Comparing performance across conditions reveals how much each type of prior knowledge contributes to temporal reasoning in current MLLMs.

4.2 Event Progress Estimation for VLA

This task evaluates a model's ability to establish a globally coherent temporal ordering of event stages under a continuous scalar scoring formulation. Given five shuffled images sampled from distinct phases of the same event, the model is required to assign each image a real-valued progress score in the range of 0% to 100%, indicating its position along the event timeline. Although the model outputs continuous scores, evaluation is performed exclusively on the temporal sequence induced by sorting these predicted scores. The resulting ordering is compared against the ground-truth chronological sequence via ranking-based consistency metrics. The numerical magnitudes of the predicted scores are not directly evaluated. Therefore, the task does not assess regression accuracy in the conventional sense, but instead measures whether the learned scoring function preserves the true temporal order of the event.

Compared with discrete permutation prediction, the continuous scoring formulation encourages the model to construct an internal scalar representation of event progression. To produce a temporally consistent ordering, the model must map visual observations onto a monotonic temporal axis that is consistent with the structural progression of the event. Consequently, the task probes the model's capacity to learn an order-preserving representation of event dynamics.

4.3 Future Frame Prediction for World Models

This task is specifically designed to evaluate World Models by assessing their capability to predict and generate the next visual state in an evolving event. The model receives four sequentially ordered images and their corresponding global description as input. Its objective is to infer the scene's visual appearance at the next moment and subsequently generate the corresponding fifth frame. The core goal is to probe the model's ability to precisely model an event's dynamic evolution, encompassing object motion trends, interaction dynamics, spatial layout continuity, and consistency in visual style.

We select 300 samples from **Orderbench** that exhibit clear visual progression and are suitable for generation-based evaluation. This task complements the previous two by shifting from temporal understanding to temporal extrapolation: the model must not only comprehend event dynamics but also synthesize a visually plausible continuation.

4.4 Metrics

For Temporal Ordering & Progress Estimation: The following metrics evaluate the accuracy of temporal structure recovery in the Multi-Image Temporal Ordering and Event Progress Estimation tasks. All metrics are normalized to $[0, 100]$, with higher values indicating better temporal ordering. Further details are provided in the Appendix **??**.

Absolute Accuracy (Acc.): correct only if the predicted order exactly matches the ground truth.

Longest Common Subsequence (L): measures partial ordering consistency via the longest order-

Table 1: **Multi-Image Temporal Ordering Results.** Performance of proprietary and open-source models on the temporal ordering task. Best results in **bold**, second-best underlined.

Model	Daily Life					Robot				
	Acc.↑	L↑	K↑	PD↑	S↑	Acc.↑	L↑	K↑	PD↑	S↑
Proprietary Models										
GPT-5	**36.85**	**79.18**	**79.14**	**71.62**	**76.65**	**21.95**	**69.72**	**65.90**	**55.99**	**63.87**
GPT-4o	26.26	74.47	73.82	64.92	71.07	6.60	61.07	56.60	42.45	53.37
Gemini-3-Flash	21.05	72.46	72.00	62.11	68.86	11.05	65.12	61.95	49.95	59.01
Claude-4-sonnet	27.55	73.42	71.77	62.82	69.33	14.60	64.36	59.71	48.82	57.63
Open-source Models										
Qwen3-VL-flash	19.50	70.59	68.89	58.56	66.01	9.50	61.77	56.30	43.97	54.01
Qwen3-VL-8B-Instruct	15.65	68.25	65.94	55.00	63.07	5.00	59.24	52.94	39.27	50.48
Qwen2.5-VL-72B-Instruct	26.50	75.73	73.38	62.73	70.61	8.00	63.24	58.53	43.72	55.16
GLM4.5V	15.37	69.07	65.96	53.61	62.88	9.86	66.36	61.77	46.94	58.36

preserving subsequence shared by prediction and ground truth.

Inversion Count (K): counts pairwise inversions relative to the ground truth, reflecting global ranking errors.

Position Deviation (PD): sums the absolute rank differences between predicted and ground-truth positions.

Overall Score (S): average of normalized L, K, and PD, serving as an aggregate measure of ordering quality.

For Next Frame Prediction: The following metrics are used for the Next Frame Prediction task, focusing on the perceptual and structural quality of the generated frames.

CLIP: cosine similarity between CLIP embeddings of predicted and ground-truth frames.

LPIPS: perceptual distance based on deep features; lower is better.

SSIM: structural similarity comparing luminance, contrast, and structure.

5 Experiments

Setting. We conduct a comprehensive evaluation of state-of-the-art models across all three task paradigms under zero-shot settings. All models are evaluated under identical data splits, shuffle schemes, and prompt templates to ensure fair comparison. For *Multi-Image Temporal Ordering*, we evaluate closed-source models including GPT-5 (Singh et al., 2025), GPT-4o (Hurst et al., 2024), Gemini-3-Flash, and Claude-4-Sonnet (Anthropic, 2025), alongside open-source models including Qwen2.5-VL-72B (Bai et al., 2025b), Qwen3-VL-8B/Flash (Bai et al., 2025a), and GLM-4.5V (GLM et al., 2024). For *Event Progress Estimation*, we additionally evaluate Gemini-Robotics-ER-1.5-Preview (Abdolmaleki et al., 2025), a vision-language-action model designed for goal-conditioned reasoning. For *Future Frame Prediction*, we evaluate the world model Emu3.5 (Cui et al., 2025) and a pipeline baseline combining GPT-4o with Wan2.5-T2I (Wan et al., 2025).

5.1 Main Results

Multi-Image Temporal Ordering. As shown in Table 1, proprietary models consistently outperform open-source models across both domains, with GPT-5 achieving the best overall performance. Nevertheless, even the strongest model remains below 37% Absolute Accuracy on Daily Life and below 22% on Robot, indicating that **multi-image temporal ordering poses a considerable challenge** for current models.

All models exhibit a noticeable performance drop on the Robot subset, with a gap of 10–17 points in overall score. This suggests that temporal reasoning becomes substantially more difficult in embodied scenarios, highlighting **limitations in cross-domain generalization**. Among open-source models, Qwen2.5-VL-72B achieves competitive results on Daily Life but suffers the largest domain gap, while GLM-4.5V maintains the most stable cross-domain performance.

Event Progress Estimation. As shown in Table 2, GPT-5 achieves the best performance across all metrics on both domains, substantially outperforming all other models including the task-specialized VLA model Gemini-Robotics-ER-1.5-Preview. This suggests that strong general-purpose reasoning can compensate for the lack of embodied training data. Notably, Gemini-Robotics-ER-1.5-Preview ranks second on Daily Life but falls behind Gemini-3-Flash on Robot, indicating that its embodied training does not consistently transfer across all robotic scenarios.

Table 2: **Event Progress Estimation Results.** Performance of MLLMs and VLA models on predicting task completion percentages from shuffled frames. Best results in **bold**, second-best underlined.

Model	Daily Life					Robot				
	Acc.↑	L↑	K↑	PD↑	S↑	Acc.↑	L↑	K↑	PD↑	S↑
Multi-modality Large Language Models										
GPT-4o	23.70	73.80	74.14	64.82	70.92	7.95	63.43	60.84	47.56	57.28
GPT-5	**36.70**	**79.62**	**80.58**	**73.20**	**77.80**	**27.00**	**75.93**	**76.16**	**67.41**	**73.16**
Gemini-3-Flash	21.15	73.27	73.97	64.17	70.47	11.65	68.07	67.15	55.69	63.64
Qwen2.5-VL-72B-Instruct	15.65	70.14	69.38	58.60	66.04	4.75	60.84	57.02	42.62	53.50
Qwen3-VL-8B-Instruct	8.25	65.37	63.41	50.62	59.80	2.90	58.96	53.87	39.04	50.62
GLM4.5V	17.22	69.45	67.80	57.07	64.77	3.43	59.18	54.50	39.76	51.15
Vision-language Action Models										
Gemini-robotics-er-1.5-preview	30.65	77.15	77.58	69.08	74.60	10.89	64.79	62.45	49.81	59.02

Table 3: **Future Frame Prediction Results.** Comparison of GPT-4o (with Wan2.5-T2I) and world model Emu3.5 on generating the fifth frame given four ordered frames. Wan2.5 is a commercially used model. Best results in **bold**.

Model	Generation Model Size	Robot			Daily Life		
		CLIP↑	LPIPS↓	SSIM↑	CLIP↑	LPIPS↓	SSIM↑
Multi-modality Large Language Models							
GPT-4o + Wan2.5-T2I	NA	0.6626	0.5478	0.0667	0.7238	0.5686	0.0893
World Models							
Emu3.5	34B	**0.8783**	0.4582	0.4077	0.8628	**0.4733**	**0.4544**

Most open-source MLLMs exhibit substantially weaker performance, with strict accuracy below 18% on Daily Life and below 5% on Robot, suggesting **limited capability in fine-grained progress estimation**. Compared with the ordering task, the cross-domain degradation is even more pronounced, further revealing **fundamental limitations in embodied reasoning and continuous temporal comprehension**.

Future Frame Prediction Table 3 compares the state-of-the-art world model Emu3.5 (Cui et al., 2025) with GPT-4o combined with Wan2.5-T2I under the same evaluation protocol. The results show that Emu3.5 achieves substantially better performance in future-frame generation.

5.2 Failure Case Analysis

To better understand the limitations of current models, we analyze representative failure cases. The observed errors reveal several recurring patterns. Some predictions violate **physical constraints and common-sense knowledge**. For example, models may confuse plant growth stages or reverse logically constrained action sequences such as jumping and landing, indicating insufficient grounding in real-world dynamics. Another common issue is **limited sensitivity to fine-grained temporal differences**. When progression gaps between closely related stages are subtle, the models often fail to capture correct temporal transitions, leading to ordering errors. In specialized scenarios, errors frequently reflect **weak domain-specific understanding**. Incorrect sequencing of irreversible steps in scientific or procedural processes suggests limited knowledge of structured event dynamics. Certain cases also exhibit **insufficient causal reasoning**, where effect events are predicted before their causes, highlighting difficulties in modeling event dependencies. Overall, these patterns suggest persistent challenges in real-world grounding, fine-grained temporal perception, and structured causal reasoning.

6 Conclusion

In this work, we introduced **Orderbench** a unified benchmark for evaluating fine-grained temporal and causal reasoning across multimodal large language models, world models, and vision–language–action systems. By formulating frame re-ranking, progress estimation, and future frame prediction within a single framework, **Orderbench** bridges perception, prediction, and embodiment under consistent evaluation principles. Extensive experiments reveal that even advanced models still struggle with temporal ordering and causal inference, highlighting a fundamental gap between static perception and dynamic understanding. We hope that **Orderbench** will serve as a foundation for developing more temporally aware and causally grounded multimodal and embodied AI models.

References

Abbas Abdolmaleki, Saminda Abeyruwan, Joshua Ainslie, Jean-Baptiste Alayrac, Montserrat Gonzalez Arenas, Ashwin Balakrishna, Nathan Batchelor, Alex Bewley, Jeff Bingham, Michael Bloesch, and 1 others. 2025. Gemini robotics 1.5: Pushing the frontier of generalist robots with advanced embodied reasoning, thinking, and motion transfer. *arXiv preprint arXiv:2510.03342*.

Anthropic. 2025. Introducing claude 4. *Preprint*.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, and 45 others. 2025a. Qwen3-vl technical report. *Preprint*, arXiv:2511.21631.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025b. Qwen2.5-vl technical report. *Preprint*, arXiv:2502.13923.

Tingzhu Bi, Yicheng Pan, Xinrui Jiang, Huize Sun, Meng Ma, and Ping Wang. 2025. Uncle: Towards scalable dynamic causal discovery in non-linear temporal systems. *arXiv preprint arXiv:2511.03168*.

Ellis Brown, Arijit Ray, Ranjay Krishna, Ross Girshick, Rob Fergus, and Saining Xie. 2025. SIMS-V: Simulated instruction-tuning for spatial video understanding. *arXiv preprint arXiv:2511.04668*.

Qingwen Bu, Jisong Cai, Li Chen, Xiuqi Cui, Yan Ding, Siyuan Feng, Shenyuan Gao, Xindong He, Xuan Hu, Xu Huang, and 1 others. 2025. Agibot world colosseo: A large-scale manipulation platform for scalable and intelligent embodied systems. *arXiv preprint arXiv:2503.06669*.

Ziming Cheng, Binrui Xu, Lisheng Gong, Zuhe Song, Tianshuo Zhou, Shiqi Zhong, Siyu Ren, Mingxiang Chen, Xiangchao Meng, Yuxin Zhang, and 1 others. 2025. Evaluating mllms with multimodal multi-image reasoning benchmark. *arXiv preprint arXiv:2506.04280*.

Suhwan Choi, Yunsung Lee, Yubeen Park, Chris Dongjoo Kim, Ranjay Krishna, Dieter Fox, and Youngjae Yu. 2026. vla-eval: A unified evaluation harness for vision-language-action models. *arXiv preprint arXiv:2603.13966*.

StarVLA Community. 2026. Starvla: A lego-like codebase for vision-language-action model developing. *arXiv preprint arXiv:2604.05014*.

Yufeng Cui, Honghao Chen, Haoge Deng, Xu Huang, Xinghang Li, Jirong Liu, Yang Liu, Zhuoyan Luo, Jinsheng Wang, Wenxuan Wang, and 1 others. 2025. Emu3. 5: Native multimodal models are world learners. *arXiv preprint arXiv:2510.26583*.

Zachary A Daniels and Dimitris Metaxas. 2022. A dynamic data driven approach for explainable scene understanding. *arXiv preprint arXiv:2206.09089*.

Hang Du, Jiayang Zhang, Guoshun Nan, Wendi Deng, Zhenyan Chen, Chenyang Zhang, Wang Xiao, Shan Huang, Yuqi Pan, Tao Qi, and 1 others. 2025. From easy to hard: The mir benchmark for progressive interleaved multi-image reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 859–869.

Mohamed Fazli Imam, Chenyang Lyu, and Alham Fikri Aji. 2025. Can multimodal llms do visual temporal understanding and reasoning? the answer is no! *arXiv e-prints*, pages arXiv–2501.

Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A Smith, Wei-Chiu Ma, and Ranjay Krishna. 2024. Blink: Multimodal large language models can see but not perceive. In *European Conference on Computer Vision*, pages 148–166. Springer.

Sara Ghazanfari, Siddharth Garg, Nicolas Flammarion, Prashanth Krishnamurthy, Farshad Khorrami, and Francesco Croce. 2024. Towards unified benchmark and models for multi-modal perceptual metrics. *arXiv preprint arXiv:2412.10594*.

Rohit Girdhar and Deva Ramanan. 2019. Cater: A diagnostic dataset for compositional actions and temporal reasoning. *arXiv preprint arXiv:1910.04744*.

Maksim Gladyshev, Natasha Alechina, Mehdi Dastani, Dragan Doder, and Brian Logan. 2025. Temporal causal reasoning with (non-recursive) structural equation models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 14949–14957.

Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, Hao Yu, Hongning Wang, Jiadai Sun, Jiajie Zhang, Jiale Cheng, Jiayi Gui, Jie Tang, Jing Zhang, Juanzi Li, and 37 others. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *Preprint*, arXiv:2406.12793.

Pranav Guruprasad, Sudipta Chowdhury, Harsh Sikka, Mridul Sharma, Helen Lu, Sean Rivera, Aryan Khurana, Hangliang Ren, and Yangyue Wang. 2025. Benchmarking the generality of vision-language-action models. *arXiv preprint arXiv:2512.11315*.

Mariya Hendriksen, Tabish Rashid, David Bignell, Raluca Georgescu, Abdelhak Lemkhenter, Katja Hofmann, Sam Devlin, and Sarah Parisot. 2025. Adapting vision-language models for evaluating world models. *arXiv preprint arXiv:2506.17967*.

Kai Hu, Jie Shao, Yuan Liu, Bhiksha Raj, Marios Savvides, and Zhiqiang Shen. 2021. Contrast and order representations for video self-supervised learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7939–7949.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Dongfu Jiang, Xuan He, Huaye Zeng, Cong Wei, Max Ku, Qian Liu, and Wenhu Chen. 2024. Mantis: Interleaved multi-image instruction tuning. *arXiv preprint arXiv:2405.01483*.

Bozheng Li, Mushui Liu, Gaoang Wang, and Yunlong Yu. 2025. Frame order matters: A temporal sequence-aware model for few-shot action recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 18218–18226.

Juncheng Li, Kaihang Pan, Zhiqi Ge, Minghe Gao, Wei Ji, Wenqiao Zhang, Tat-Seng Chua, Siliang Tang, Hanwang Zhang, and Yueting Zhuang. 2023. Fine-tuning multimodal llms to follow zero-shot demonstrative instructions. *arXiv preprint arXiv:2308.04152*.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, and 1 others. 2024a. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22195–22206.

Shicheng Li, Lei Li, Yi Liu, Shuhuai Ren, Yuanxin Liu, Rundong Gao, Xu Sun, and Lu Hou. 2024b. Vitatecs: A diagnostic dataset for temporal concept understanding of video-language models. In *European Conference on Computer Vision*, pages 331–348. Springer.

Yu-Chien Liao, Hsi-Che Lin, Yu-Chu Yu, Yen-Chun Chen, Yu-Chiang Frank Wang, and 1 others. 2024. Rextime: A benchmark suite for reasoning-across-time in videos. *arXiv e-prints*, pages arXiv–2406.

Shanchuan Lin, Linjie Yang, Imran Saleemi, and Soumyadip Sengupta. 2022. Robust high-resolution video matting with temporal guidance. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 238–247.

Pei Liu, Qingtian Ning, Xinyan Lu, Haipeng Liu, Weiliang Ma, Dangen She, Peng Jia, Xianpeng Lang, and Jun Ma. 2025. Omnireason: A temporal-guided vision-language-action framework for autonomous driving. *arXiv preprint arXiv:2509.00789*.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, and 1 others. 2024. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pages 216–233. Springer.

Yueen Ma, Zixing Song, Yuzheng Zhuang, Jianye Hao, and Irwin King. 2024. A survey on vision-language-action models for embodied ai. *arXiv preprint arXiv:2405.14093*.

Ajay Mandlekar, Soroush Nasiriany, Bowen Wen, Iretiayo Akinola, Yashraj Narang, Linxi Fan, Yuke Zhu, and Dieter Fox. 2023. Mimicgen: A data generation system for scalable robot learning using human demonstrations. *arXiv preprint arXiv:2310.17596*.

Abby O'Neill, Abdul Rehman, Abhiram Maddukuri, Abhishek Gupta, Abhishek Padalkar, Abraham Lee, Acorn Pooley, Agrim Gupta, Ajay Mandlekar, Ajinkya Jain, and 1 others. 2024. Open x-embodiment: Robotic learning datasets and rt-x models: Open x-embodiment collaboration 0. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6892–6903. IEEE.

Alkesh Patel, Joel Ruben Antony Moniz, Roman Nguyen, Nick Tzou, Hadas Kotek, and Vincent Renkens. 2021. Mmiu: Dataset for visual intent understanding in multimodal assistants. *arXiv preprint arXiv:2110.06416*.

Ziyao Shangguan, Chuhan Li, Yuxuan Ding, Yanan Zheng, Yilun Zhao, Tesca Fitzgerald, and Arman Cohan. 2024. Tomato: Assessing visual temporal reasoning capabilities in multimodal foundation models. *arXiv preprint arXiv:2410.23266*.

Prafull Sharma, Ayush Tewari, Yilun Du, Sergey Zakharov, Rares Ambrus, Adrien Gaidon, William T Freeman, Frédo Durand, Joshua B Tenenbaum, and Vincent Sitzmann. 2022. Seeing 3d objects in a single image via self-supervised static-dynamic disentanglement. *arXiv preprint*.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, and 1 others. 2025. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*.

Yingjin Song, Yupei Du, Denis Paperno, and Albert Gatt. 2025. Burn after reading: Do multimodal large language models truly capture order of events in image sequences? *arXiv preprint arXiv:2506.10415*.

Jingqi Tong, Yurong Mou, Hangcheng Li, Mingzhe Li, Yongzhuo Yang, Ming Zhang, Qiguang Chen, Tianyi Liang, Xiaomeng Hu, Yining Zheng, Xinchi Chen, Jun Zhao, Xuanjing Huang, and Xipeng Qiu. 2025. Thinking with video: Video generation as a promising multimodal reasoning paradigm. *Preprint*, arXiv:2511.04570.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, and 43 others. 2025. Wan: Open and advanced large-scale video generative models. *Preprint*, arXiv:2503.20314.

Fei Wang, Xingyu Fu, James Y Huang, Zekun Li, Qin Liu, Xiaogeng Liu, Mingyu Derek Ma, Nan Xu, Wenxuan Zhou, Kai Zhang, and 1 others. 2024a. Muirbench: A comprehensive benchmark for robust multi-image understanding. *arXiv preprint arXiv:2406.09411*.

Xiaochen Wang, Heming Xia, Jialin Song, Longyu Guan, Qingxiu Dong, Rui Li, Yixin Yang, Yifan Pu, Weiyao Luo, Yiru Wang, and 1 others. 2025a. Beyond single frames: Can lmms comprehend implicit narratives in comic strip? In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 6436–6452.

Xiaochen Wang, Heming Xia, Jialin Song, Longyu Guan, Yixin Yang, Qingxiu Dong, Weiyao Luo, Yifan Pu, Yiru Wang, Xiangdi Meng, Wenjie Li, and Zhifang Sui. 2025b. Beyond single frames: Can lmms comprehend temporal and contextual narratives in image sequences? *Preprint*, arXiv:2502.13925.

Xiyao Wang, Yuhang Zhou, Xiaoyu Liu, Hongjin Lu, Yuancheng Xu, Feihong He, Jaehong Yoon, Taixi Lu, Gedas Bertasius, Mohit Bansal, and 1 others. 2024b. Mementos: A comprehensive benchmark for multimodal large language model reasoning over image sequences. *arXiv preprint arXiv:2401.10529*.

Charig Yang, Weidi Xie, and Andrew Zisserman. 2024. Made to order: Discovering monotonic temporal changes via self-supervised video ordering. In *European Conference on Computer Vision*, pages 268–286. Springer.

Xiaojun Ye, Guanbao Liang, Chun Wang, Liangcheng Li, Pengfei Ke, Rui Wang, Bingxin Jia, Gang Huang, Qiao Sun, and Sheng Zhou. M4bench: A benchmark of multi-domain multi-granularity multi-image understanding for multi-modal large language models.

Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B Tenenbaum. 2019. Clevrer: Collision events for video representation and reasoning. *arXiv preprint arXiv:1910.01442*.

Hong Qing Yu. 2025. T-cpdl: A temporal causal probabilistic description logic for developing logic-rag agent. *arXiv preprint arXiv:2506.18559*.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, and 1 others. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567.

Yunhang Shen Yulei Qin Mengdan Zhang, Xu Lin Jinrui Yang Xiawu Zheng, Ke Li Xing Sun Yunsheng Wu, Rongrong Ji Chaoyou Fu, and Peixian Chen. 2021. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 18.

Zicheng Zhang, Haoning Wu, Erli Zhang, Guangtao Zhai, and Weisi Lin. 2024. Q-bench: A benchmark for multi-modal foundation models on low-level vision from single images to pairs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

A Dataset Details

A.1 Data Source

Daily Life Domain. All videos in the daily life domain are sourced exclusively from major online video platforms: Bilibili and YouTube.

Robot Domain. The robotics domain videos are heterogeneous, comprising two parts: one is collected from online video platforms (including Bilibili and YouTube), and the other is sourced from multiple robotic operation datasets, including both real-world and simulated environments. Details of these datasets are summarized in Table 4.

A.2 Human Annotation Guideline

Annotators shall construct a five-frame image sequence from each video and provide an accurate overall description according to the following guidelines:

- **Unambiguous sequence.** The selected frames should be ordered clearly, so that their temporal progression is evident and cannot be misinterpreted.
- **Meaningful transitions.** Consecutive frames should reflect noticeable changes in action, motion, or object state, ensuring the sequence conveys a coherent progression.
- **Frame quality.** Only select frames that are visually clear, without occlusions, or other artifacts that could hinder interpretation.
- **Description Accuracy.** The textual description must accurately reflect the event or action represented without ambiguity.

A.3 Frame Count Ablation

To determine the optimal sequence length, we conducted an ablation study over frame counts from 3 to 7 using Qwen2.5-VL-7B-Instruct on 500 samples. We jointly consider two criteria: information entropy, which reflects how uniformly model scores are distributed and thus how well the benchmark discriminates among models; and above-chance performance, which ensures the task remains solvable rather than degenerating into random guessing.

Table 4: Robotics-Related Datasets and Their Links

Dataset Name	Dataset Name (cont.)
AgiBot	G1_ToastedBread_Dataset
Berkeley Autolab UR5	Dobb·E
Berkeley Cable Routing	MimicGen
Columbia PushT Dataset	MobileALOHA
FurnitureBench	NYU VINN
G1_BlockStacking_Dataset	RoboCasa
G1_CameraPackaging_Dataset	Stanford HYDRA
G1_DualArmGrasping_Dataset	Tokyo PR2 Fridge Opening
G1_MountCamera_Dataset	TOTO Benchmarch
G1_ObjectPlacement_Dataset	Z1_DualArm_FoldClothes_Dataset
G1_Pouring_Dataset	Z1_DualArm_PourCoffee_Dataset
Z1_StackBox_Dataset	Z1_DualArmStackBox_Dataset
G1_MountCameraRedGripper_Dataset	

Table 5: Field Description of JSON Data

Field	Definition	Required
id	Unique identifier for the data sample	Yes
type	Primary category	Yes
subtype	Fine-grained subcategory	Yes
global_annotation	Text description of the entire image sequence	Yes
image_paths_main	File path list of the main view with the most complete scene information	Yes
image_paths_left	File path list of the view from the left side	Opt
image_paths_right	File path list of the view from the right side	Opt
image_paths_other	File path list of the view from the other side	Opt
selected_images	File path list of selected images integrating multi-view information for sorting	Yes
shuffled_images	File path list of shuffled selected images	Yes
order	Original correct index sequence for shuffled selected images	Yes
block_images	File path list of partially ordered sub-block images	Opt

Table 6 summarizes the key results. Three-frame sequences yield high discriminability but low entropy, as the limited combinatorial space concentrates predictions into few score bins. Six- and seven-frame sequences achieve high entropy but model accuracy drops to or below chance level, rendering the benchmark uninformative. Five frames achieves the highest normalized entropy (0.8626) while maintaining a meaningful above-chance lift of 3.1×, representing the best trade-off between task difficulty and evaluation informativeness.

The marginal difficulty analysis further supports this choice. The transition from 4 to 5 frames introduces a moderate accuracy drop of 5.4 percentage points and an entropy gain of 0.55 bits, while the transition from 5 to 6 frames yields diminishing entropy returns (-0.02 bits) alongside a sharp decline in model performance that pushes accuracy to near-chance levels. We therefore adopt five frames as the standard sequence length for **Orderbench**.

A.4 Annotation Consistency

To verify that the temporal orderings in **Orderbench** are unambiguous and reproducible, we assessed inter-annotator agreement on a randomly sampled subset. Four independent annotators, distinct from the original annotation team, re-sorted each sequence without access to the ground-truth ordering.

Table 7 reports per-annotator Kendall's τ against the ground truth. The mean inter-annotator τ is 0.7917, and the mean annotator-vs-ground-truth τ is 0.8678, confirming strong agreement. Annotators C and B achieve near-perfect consistency ($\tau > 0.92$), while Annotator D shows relatively lower agreement ($\tau = 0.75$), likely due to ambiguous cases in specialized domains. Overall, these results validate that the established orderings reflect genuine temporal structure rather than subjective interpretation.

A.5 Data Structure Description

This section specifies our data format, explaining each field in the JSON structure and providing a

Table 6: Frame count ablation. $n!$ denotes the number of possible orderings. Lift indicates how many times model accuracy exceeds random chance. H_{norm} is Shannon entropy normalized by the maximum entropy of 20 bins.

Frames	$n!$	Random	Model Acc.	Lift	H_{norm}
3	6	16.67	22.20	1.3×	0.4499
4	24	4.17	8.00	1.9×	0.7342
5	**120**	**0.83**	**2.60**	**3.1×**	**0.8626**
6	720	0.14	0.20	1.4×	0.8573
7	5040	0.02	0.00	0.0×	0.8455

Table 7: Per-annotator Kendall's τ against ground truth.

Annotator	Kendall's τ vs. GT
A	0.8424
B	0.9228
C	0.9555
D	0.7506
Mean (inter-annotator)	0.7917
Mean (vs. GT)	0.8678

sample.

The detailed descriptions of the JSON fields are provided in Table 5.

The complete raw data sample is shown below:


```
{
   "id": "VideoRank_0001",
   "type": "Sports",
   "subtype": "basketball",
   "global_annotation": "The man in
    the blue jersey jumps up to
    dunk the ball into the hoop.",
   "image_paths": [
     "VideoRank_0001/frame_0001.jpg",
     "VideoRank_0001/frame_0002.jpg",
     "VideoRank_0001/frame_0003.jpg",
     "VideoRank_0001/frame_0004.jpg",
     "VideoRank_0001/frame_0005.jpg"
   ],
   "image_paths_left": [],
   "image_paths_right": [],
   "image_paths_other": []
 }
```


The complete shuffle data sample is shown below:


```
{
    "id": "VideoRank_0001",
    "type": "Sports",
    "subtype": "basketball",
    "global_annotation": "The man in the
     blue jersey jumps up to
     dunk the ball into the hoop.",
    "selected_images": [
      "VideoRank_0001/frame_0001.jpg",
      "VideoRank_0001/frame_0002.jpg",
      "VideoRank_0001/frame_0003.jpg",
      "VideoRank_0001/frame_0004.jpg",
      "VideoRank_0001/frame_0005.jpg"
    ],
    "shuffled_images": [
      "VideoRank_0001/frame_0005.jpg",
      "VideoRank_0001/frame_0004.jpg",
      "VideoRank_0001/frame_0001.jpg",
      "VideoRank_0001/frame_0002.jpg",
      "VideoRank_0001/frame_0003.jpg"
    ],
    "order": [3,4,5,2,1],
    "block_images": [
      "VideoRank_0001/frame_0003.jpg",
      "VideoRank_0001/frame_0004.jpg"
    ]
}
```


A.6 Human Performance

To establish a reference for benchmark difficulty and to better contextualize model performance, we evaluated human performance on the proposed task to quantify the gap between models and humans.Given the impracticality of manually evaluating the entire benchmark, we employed a stratified random sampling strategy, selecting three instances per class across the two domains, Daily Life and Robotics, thereby ensuring diversity.Five undergraduate evaluators were recruited to perform the task on the sampled subsets. Each evaluator received detailed instructions and completed a brief practice session prior to the formal evaluation.

The resulting human performance on these subsets, presented in Table 8 and Table 9, provides a reference for the benchmark, revealing a substantial performance gap between models and humans.

A.7 Evaluation Metrics

The following are detailed explanations of several metrics.

Table 8: Human Performance on Multi-Image Temporal Ordering(With Global Description Condition)

Setting	Daily Life					Robot				
	Acc.↑	L↑	K↑	PD↑	S↑	Acc.↑	L↑	K↑	PD↑	S↑
Human	70.72	90.86	90.98	87.66	89.83	74.79	92.54	92.93	90.16	91.88
GPT-4o	28.20	75.97	76.61	68.19	73.59	8.45	63.94	61.29	48.00	57.74

Table 9: Human Performance on Multi-Image Temporal Ordering(With Partial Ordered Block Condition)

Setting	Daily Life					Robot				
	Acc.↑	L↑	K↑	PD↑	S↑	Acc.↑	L↑	K↑	PD↑	S↑
Human	85.96	95.72	95.64	93.85	95.07	83.62	94.96	95.18	93.36	94.50
Gemini-2.5-Flash	37.25	81.08	80.29	71.71	77.70	14.16	67.76	63.40	49.88	60.35

- **Absolute Accuracy (Acc.).** This metric measures strict sequence-level correctness, where a prediction is considered correct only if the predicted order exactly matches the ground-truth order. It evaluates whether the model can recover the full temporal permutation without any error. The metric is binary at the sample level and is reported as a percentage over the dataset.

- **Longest Common Subsequence (LCS, L).** This metric evaluates partial ordering consistency by computing the length of the longest subsequence that is shared between the predicted and ground-truth sequences while preserving order. It reflects how much of the correct temporal structure is maintained even when the full sequence is not perfectly recovered. For sequences of length $N = 5$, the normalized score is defined as

$$L_{norm} = \frac{L}{5} \times 100, \tag{1}$$

where $L \in [0, 5]$.

- **Inversion Count (K).** This metric quantifies global ranking errors by counting the number of pairwise inversions in the predicted sequence relative to the ground truth. An inversion occurs when a pair of elements appears in the incorrect relative order. It directly measures the degree of discordance between two rankings. For $N = 5$, the number of possible pairs is $52 = 10$, and the normalized score is defined as

$$K_{norm} = \left(1 - \frac{K}{10}\right) \times 100, \tag{2}$$

where $K \in [0, 10]$.

- **Position Deviation (PD).** This metric measures the overall positional displacement of elements between the predicted and ground-truth rankings. It computes the total absolute difference in rank positions across all elements, capturing the magnitude of ordering errors. Let $\hat{p}_i$ and p_i denote the predicted and ground-truth positions of element i, respectively. The normalized score is defined as

$$PD_{norm} = 100 \left(1 - \frac{1}{12} \sum_{i=1}^{5} |\hat{p}_i - p_i|\right), \tag{3}$$

where 12 is the maximum possible positional deviation for sequences of length five.

- **Overall Score (S).** The final performance score is computed as the average of the three normalized ranking metrics:

$$S = \frac{L_{norm} + K_{norm} + PD_{norm}}{3}. \tag{4}$$

A.8 More Data Visualization

To provide a more detailed understanding of our dataset, we include additional visualizations for both the daily-life and robotics domains.

Figure 4 presents representative samples from the daily portion of the dataset, illustrating the diversity of daily activities. Figure 5 presents examples from the robotic portion of our dataset, including observations from multiple viewpoints, tasks across diverse scenarios, and various manipulation contexts, highlighting the diversity of the data.

B Further Analysis of Experimental Results

B.1 Details of the Experimental Results

This subsection aims to systematically present the accuracy performance of various models in temporal reasoning tasks through comprehensive quantitative data, providing empirical evidence for subsequent in-depth analysis.

Figure 6 and Figure 7 present the subcategory-level results of GPT-4o, Gemini-2.5-Flash, GLM-4.5V, and Qwen2.5-VL-72B on the Multi-Image Temporal Ordering (with Partial Ordered Block) task in the Daily Life and Robot domains, respectively. Figure 8 and Figure 9 show the performance of GPT-4o, Gemini-2.5-Flash, GLM-4.5V, and Gemini-robotics-er-1.5-preview on the Event Progress Estimation task across different subcategories in the Daily Life and Robot domains. Figure 10 and Figure 11 further report the subcategory-level performance of Qwen3-VL-8B on the Multi-Image Temporal Ordering task under two reasoning strategies—Pairwise Relational Modeling and Experience-Guided Reasoning—evaluated separately in the Daily Life and Robot domains.

B.2 Failure Case

To gain deeper insights into the typical failure patterns of the models across different tasks, this section provides a qualitative analysis of model errors through visualized case studies. Figure 12 presents representative failure cases from the models in Multi-Image Temporal Ordering task, while Figure 13 further illustrates typical errors made by each model in the event progression estimation task.

Detailed analysis and discussion of these cases will be systematically presented in B.3.

B.3 Analysis

Based on the bar charts in Figures 6, 7, 8, 9, 10 and 11, which provide a more detailed breakdown of model accuracy on the Multi-Image Temporal Ordering and Event Progress Estimation tasks, several observations can be made.

The model performs significantly better in the daily life domain than in the robotics domain, likely due to a lack of exposure to relevant scenarios during training. Additionally, the model achieves higher accuracy on tasks with clear procedural steps and longer time spans, such as the high jump or chemical experiments. The influence of contextual information also varies across task types, with the Partial Ordered Block condition having the most pronounced effect.

Additionally, in previous evaluations on multi-image temporal sequencing and event progression estimation tasks, we observed that even state-of-the-art models still exhibit significant performance shortcomings. We present several typical failure cases in Figure 12 and 13. Through in-depth analysis of these examples, the following issues can be observed:

1) Violation of Global Event Descriptions: Models often fail to adequately understand or adhere to the global event annotations provided in the prompts, sometimes even generating sequences completely opposite to the described events. For example, in depicting a lunar eclipse, a model may incorrectly sequence the moon being obscured as a process of gradual revelation.

2) Disregard for Local Ordinal Constraints: Models frequently fail to strictly follow the partially ordered block constraints specified in the prompts. Even when explicitly instructed that certain image blocks should be treated as contiguous sequence units, models often fragment them in their outputs, disrupting local temporal coherence.

3) Contradiction with World Knowledge and Physical Laws: The sequencing results often violate common sense and fundamental physical principles. For instance, in a plant growth sequence, models might confuse the order of the "sprouting" and "leafing" stages. In a high jump sequence, they might invert the necessary order between "clearing the bar" and "landing on the mat".

4) Weak Perception of Fine-Grained Temporal Intervals: When estimating event progression, models struggle to precisely distinguish subtle progression differences between consecutive actions or states that occur within very short time intervals, leading to subsequent errors in temporal ordering.

5) Deficiency in Specific Domain Knowledge (e.g., chemical reactions, physical experiments, astronomical phenomena): Models lack a deep understanding of processes in specialized domains, resulting in incorrect sequencing of irreversible steps or phenomena.

6) Inadequate Causal Reasoning Ability: Models perform poorly in comprehending causal relationships such as "Action A leads to Outcome B." Failure cases frequently exhibit causal inversion, where the model incorrectly places the resultant event before the causal event.

C Task Details

C.1 Representative Case

Representative cases from the core tasks are presented in this subsection, demonstrating typical successful performances of the models across different temporal reasoning tasks. These examples complement the failure case analysis in Section B, together providing a comprehensive understanding of model capabilities.Figures 14, 15, 16, and 17 illustrate representative cases from the following four tasks respectively: 1) Multi-image temporal ordering with global description conditioning, 2) Multi-image temporal ordering with partial ordered block conditioning, 3) Event progress estimation, and 4) Future frame prediction.

C.2 Prompt Formats

The specific prompts for each task are provided in Figures 18, 19, and 20.

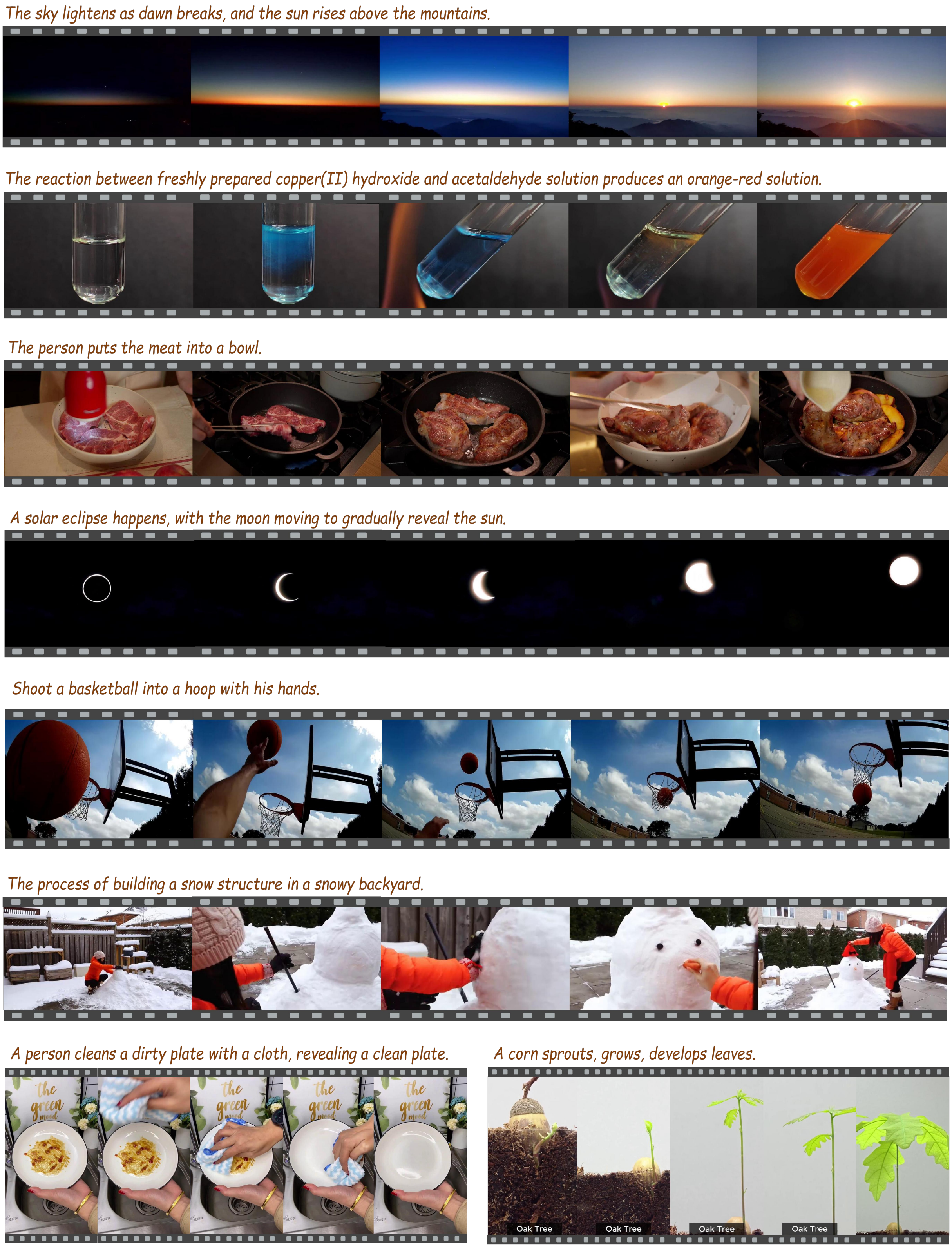

Figure 4: Some examples from the daily life portion of the dataset.

Figure 5: Some examples from the robotic portion of the dataset.

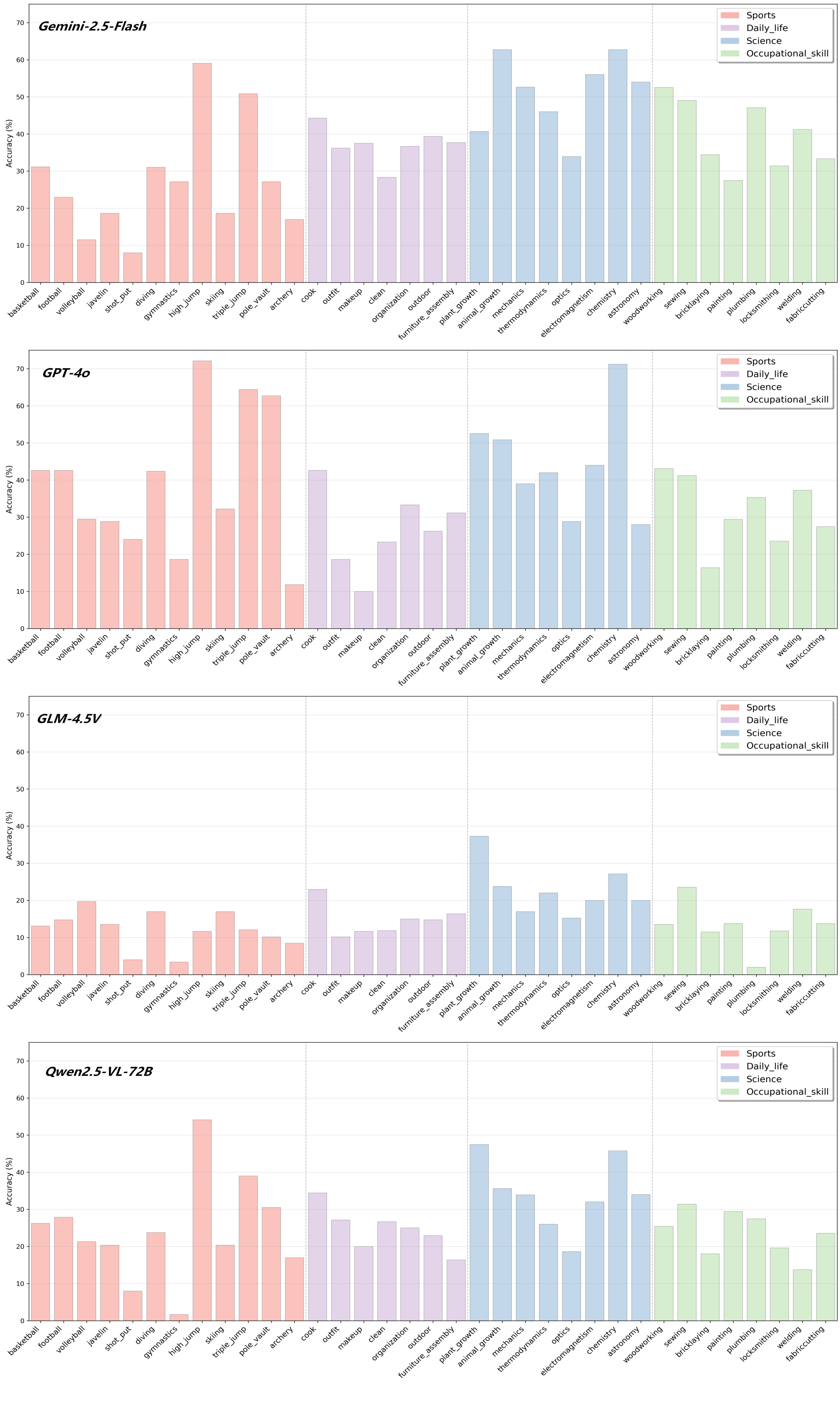

Figure 6: Per-subtype accuracy of GPT-4o, Gemini-2.5-Flash, GLM-4.5V, and Qwen2.5-VL-72B on the Multi-Image Temporal Ordering (with Partial Ordered Block) task in the Daily Life domain.

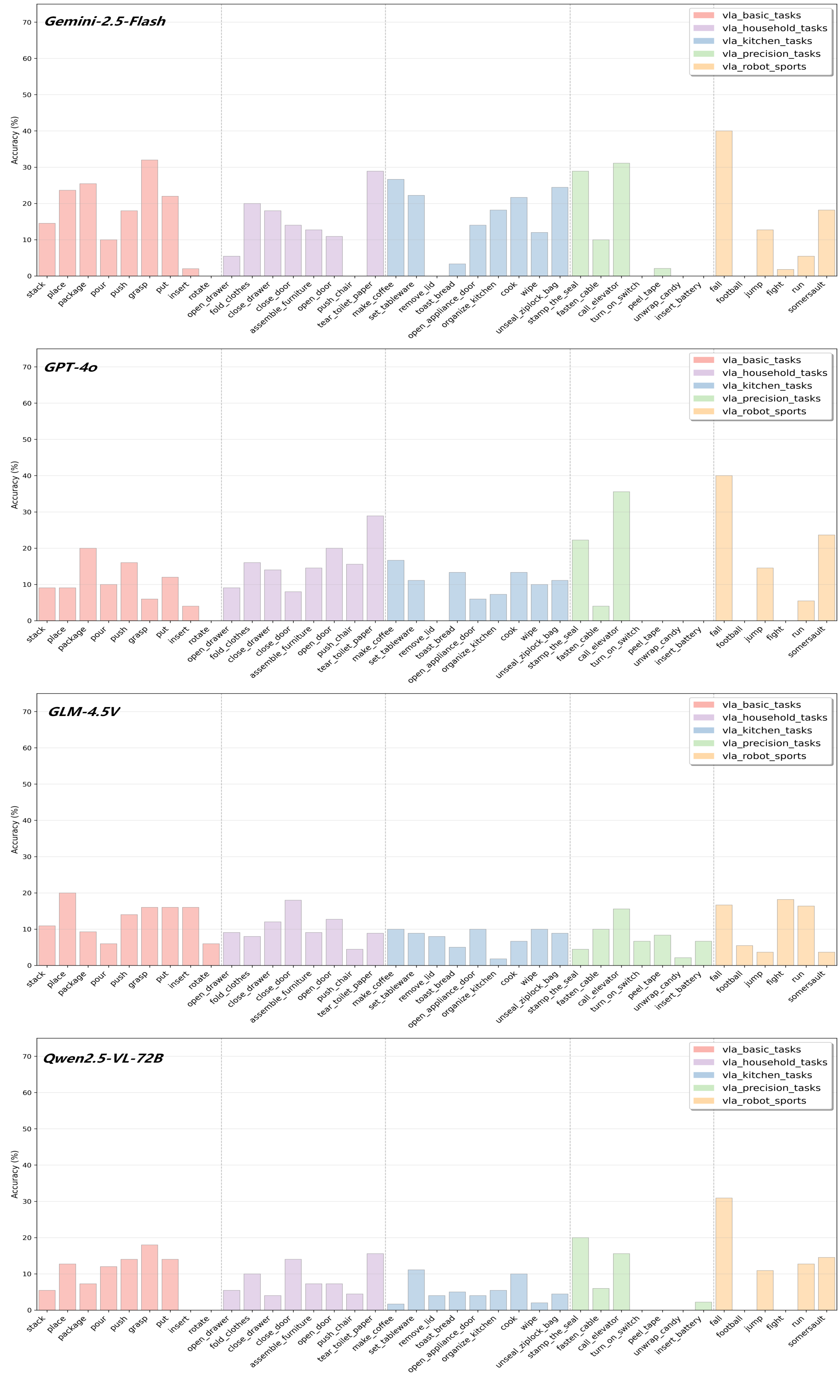

Figure 7: Per-subtype accuracy of GPT-4o, Gemini-2.5-Flash, GLM-4.5V, and Qwen2.5-VL-72B on the Multi-Image Temporal Ordering (with Partial Ordered Block) task in the Robot domain.

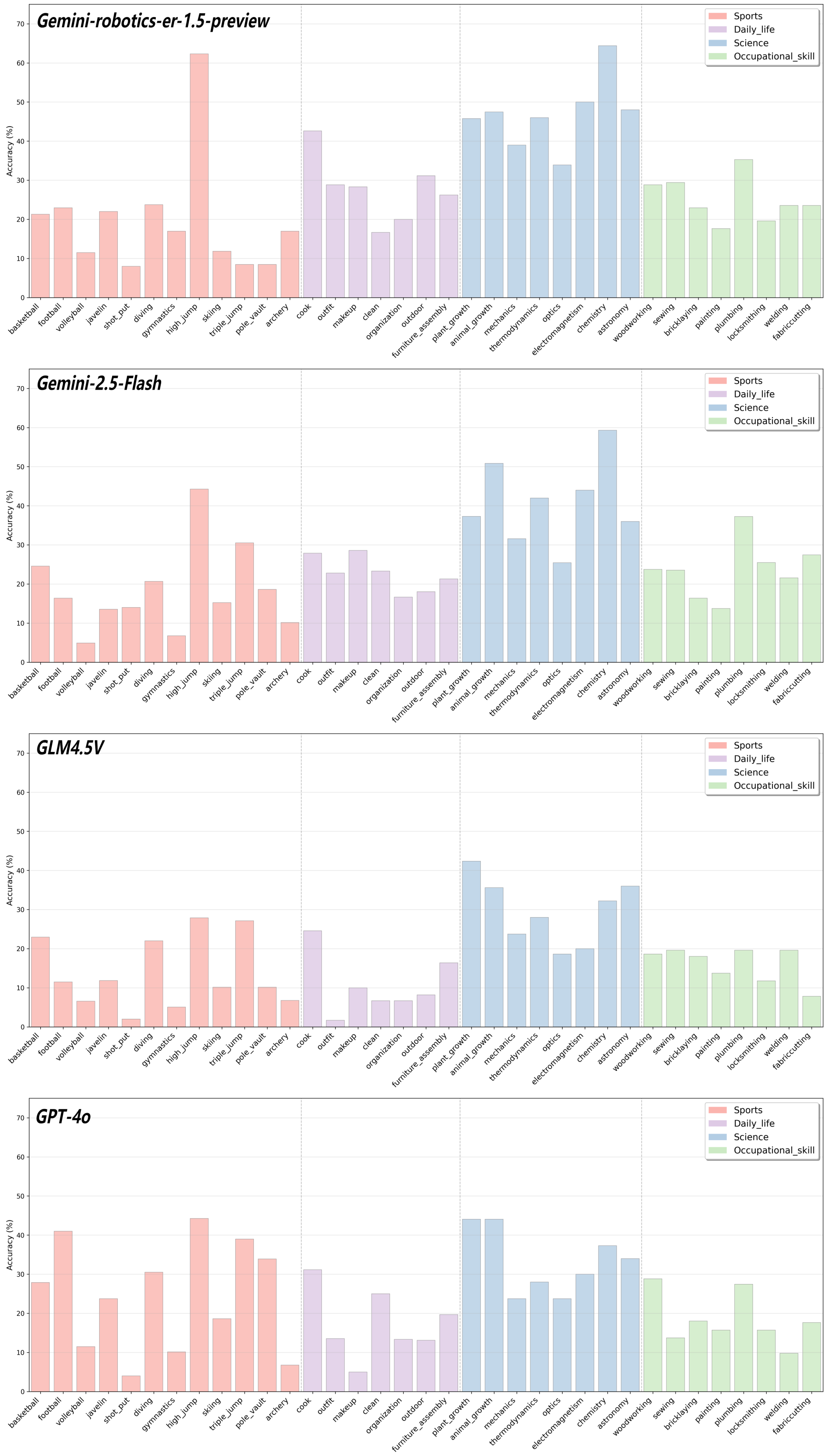

Figure 8: Per-subtype accuracy of GPT-4o, Gemini-2.5-Flash, GLM-4.5V, and Gemini-robotics-er-1.5-preview on the Event Progress Estimation task in the Daily Life domain.

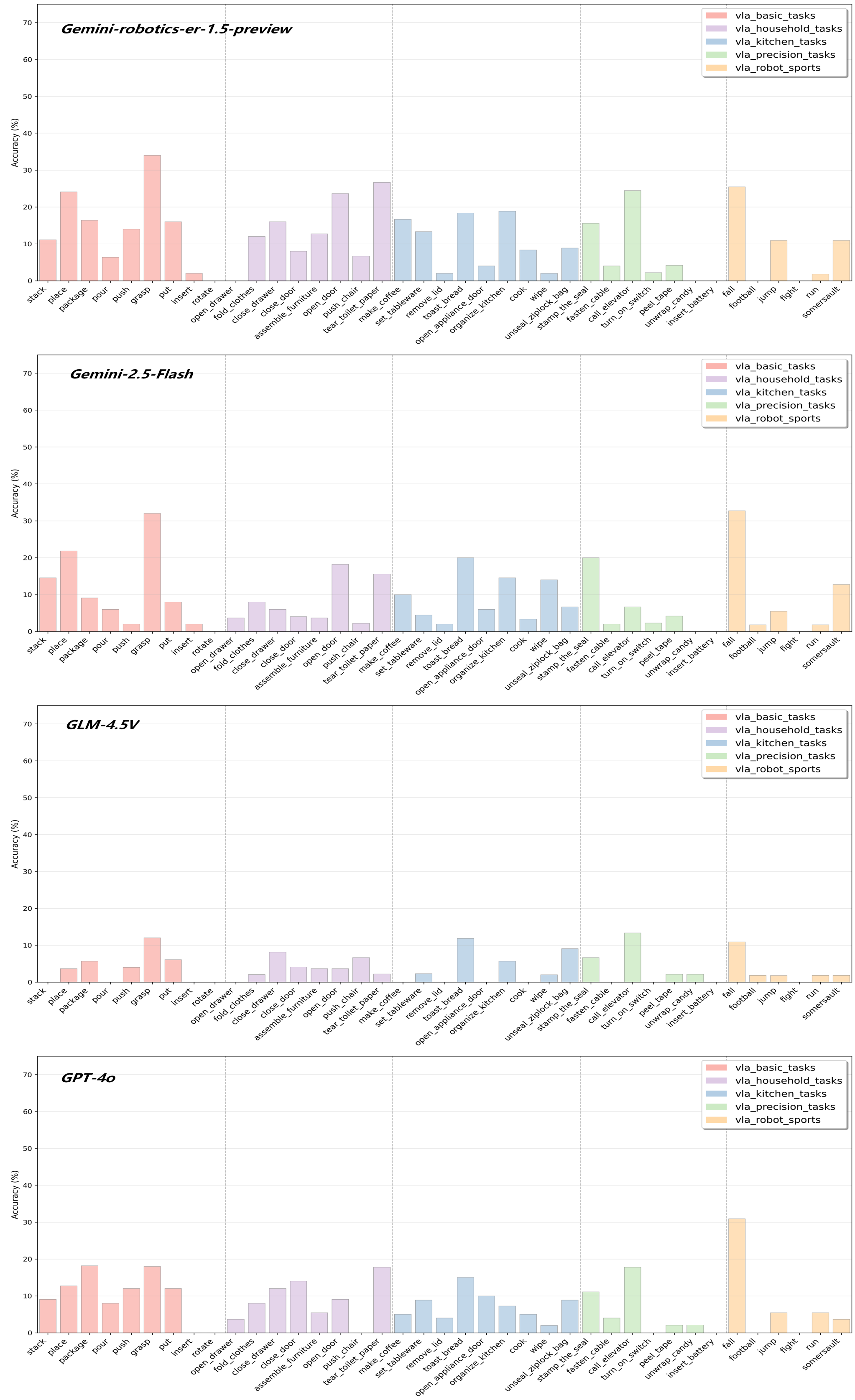

Figure 9: Per-subtype accuracy of GPT-4o, Gemini-2.5-Flash, GLM-4.5V, and Gemini-robotics-er-1.5-preview on the Event Progress Estimation task in the Robot domain.

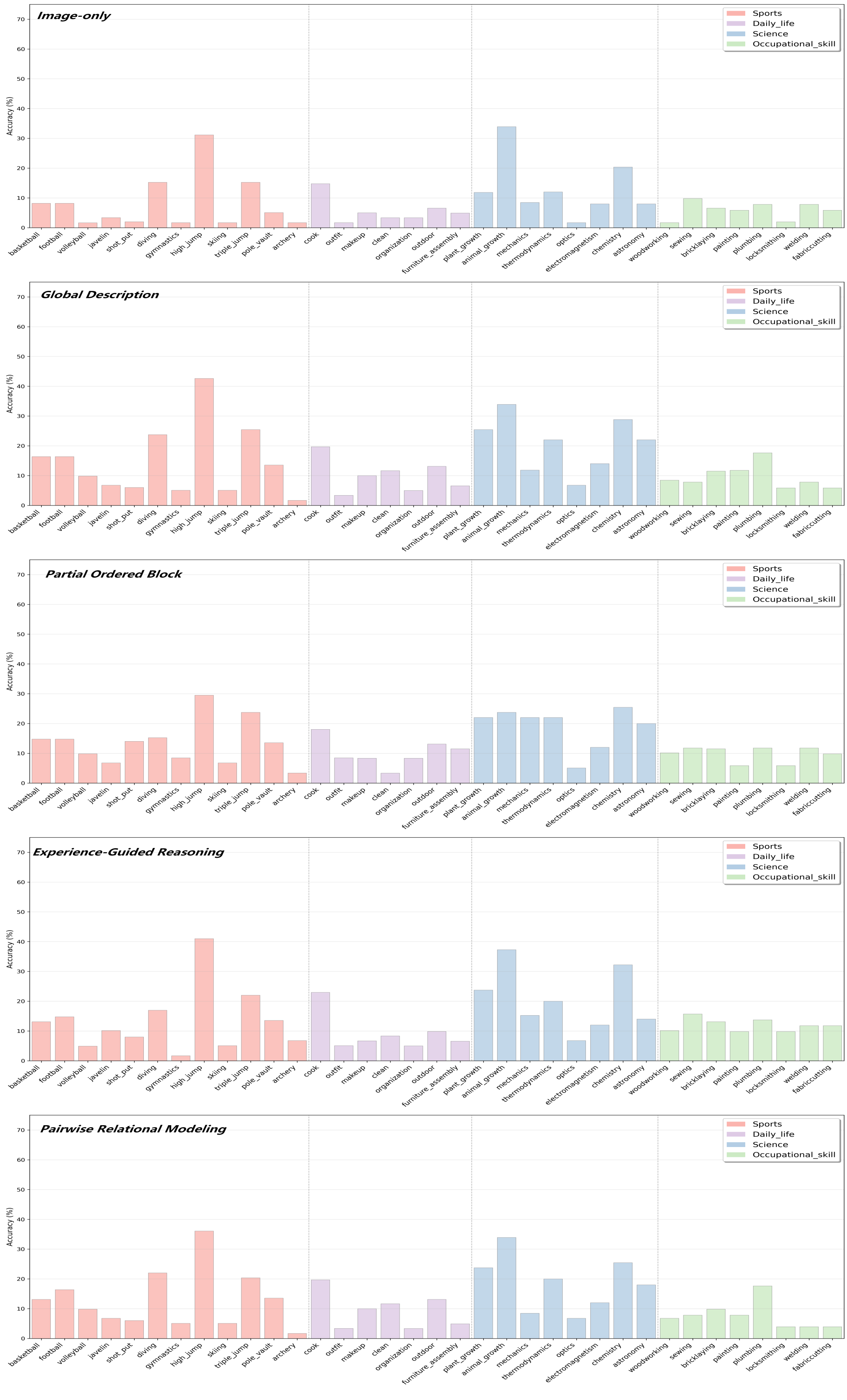

Figure 10: Per-subtype accuracy of Qwen3-VL-8B on the original Multi-Image Temporal Ordering task and its variants under Pairwise Relational Modeling and Experience-Guided Reasoning in the Daily Life domain.

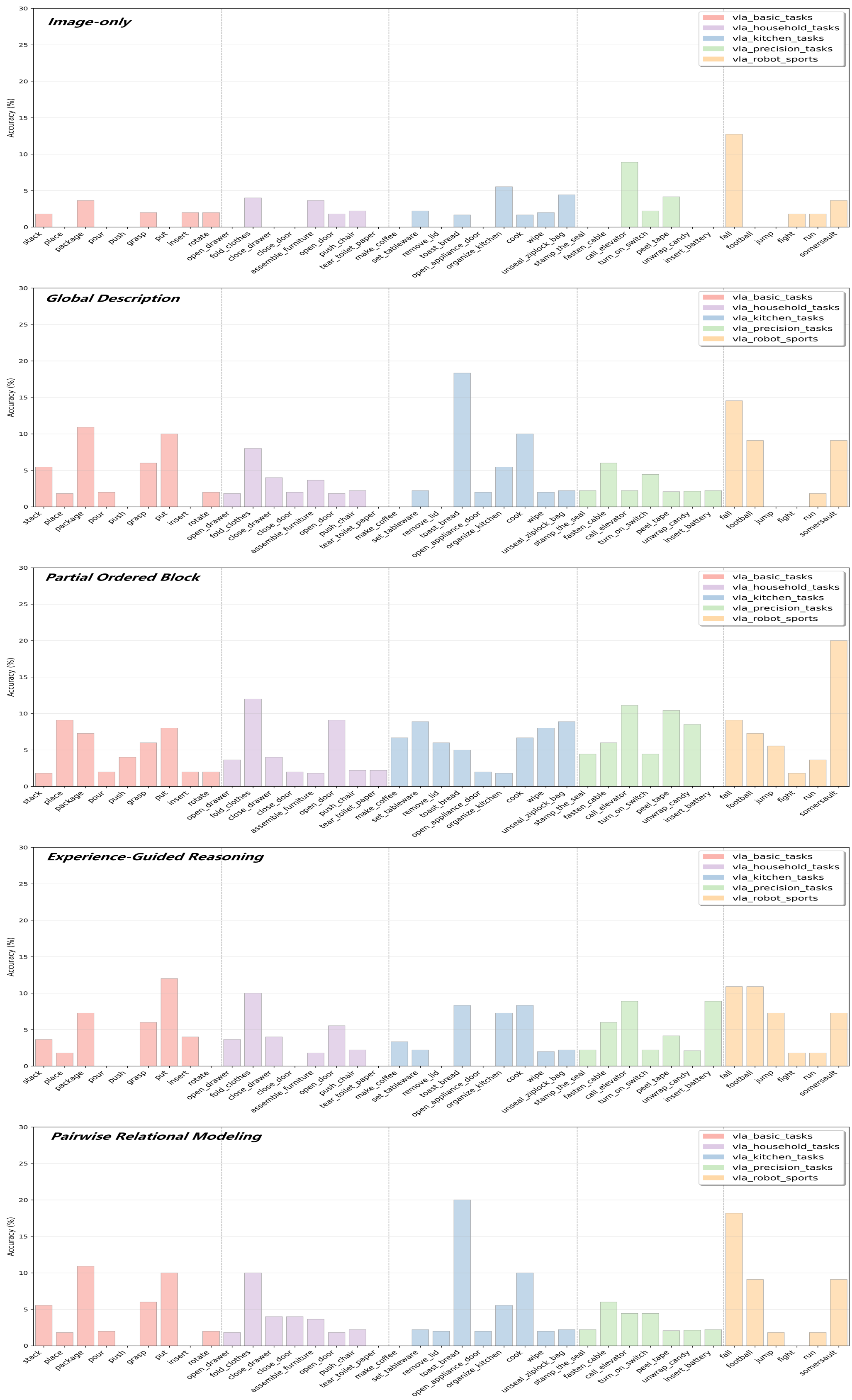

Figure 11: Per-subtype accuracy of Qwen3-VL-8B on the original Multi-Image Temporal Ordering task and its variants under Pairwise Relational Modeling and Experience-Guided Reasoning in the Robot domain.

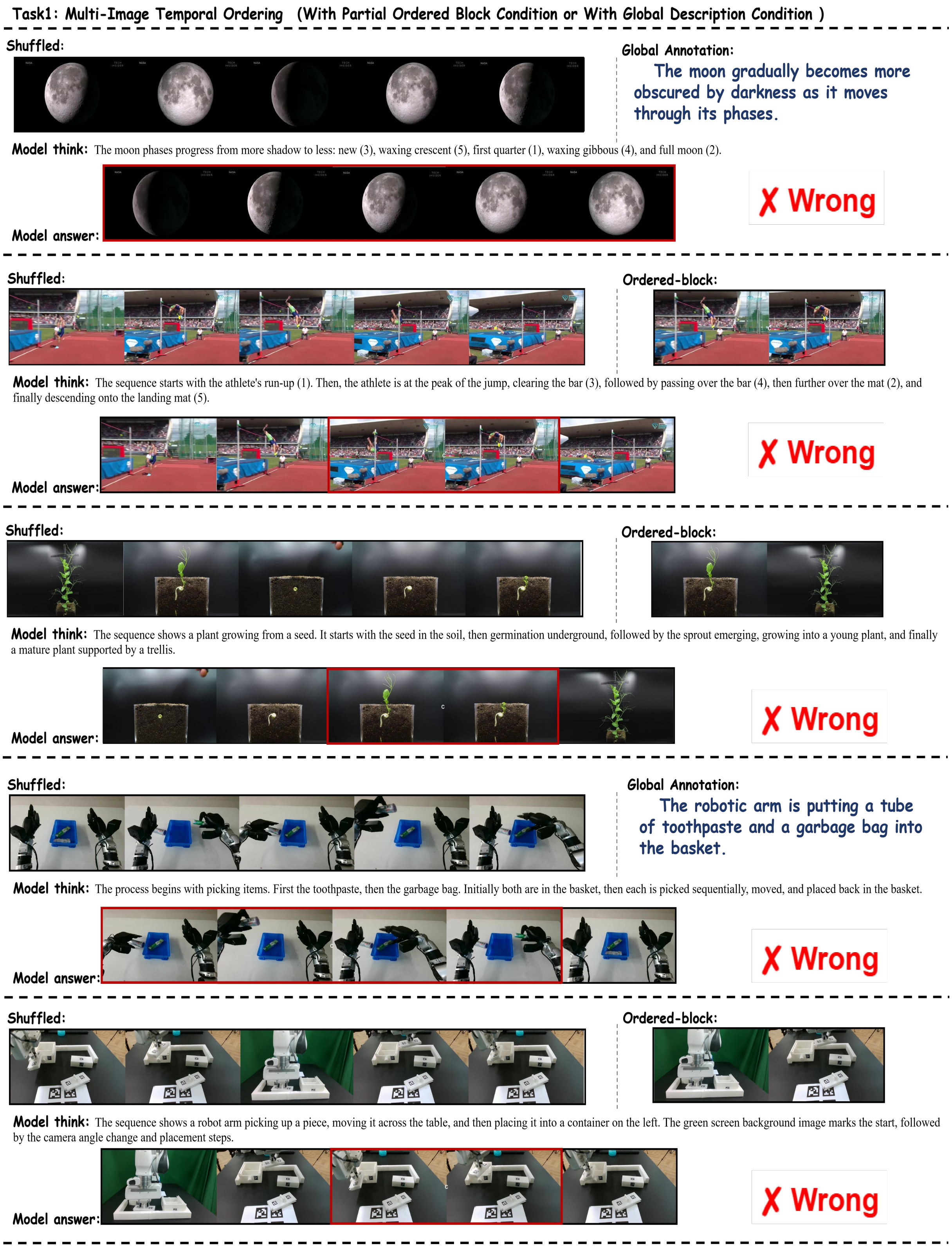

Figure 12: Failure cases in the Multi-Image Temporal Ordering task.

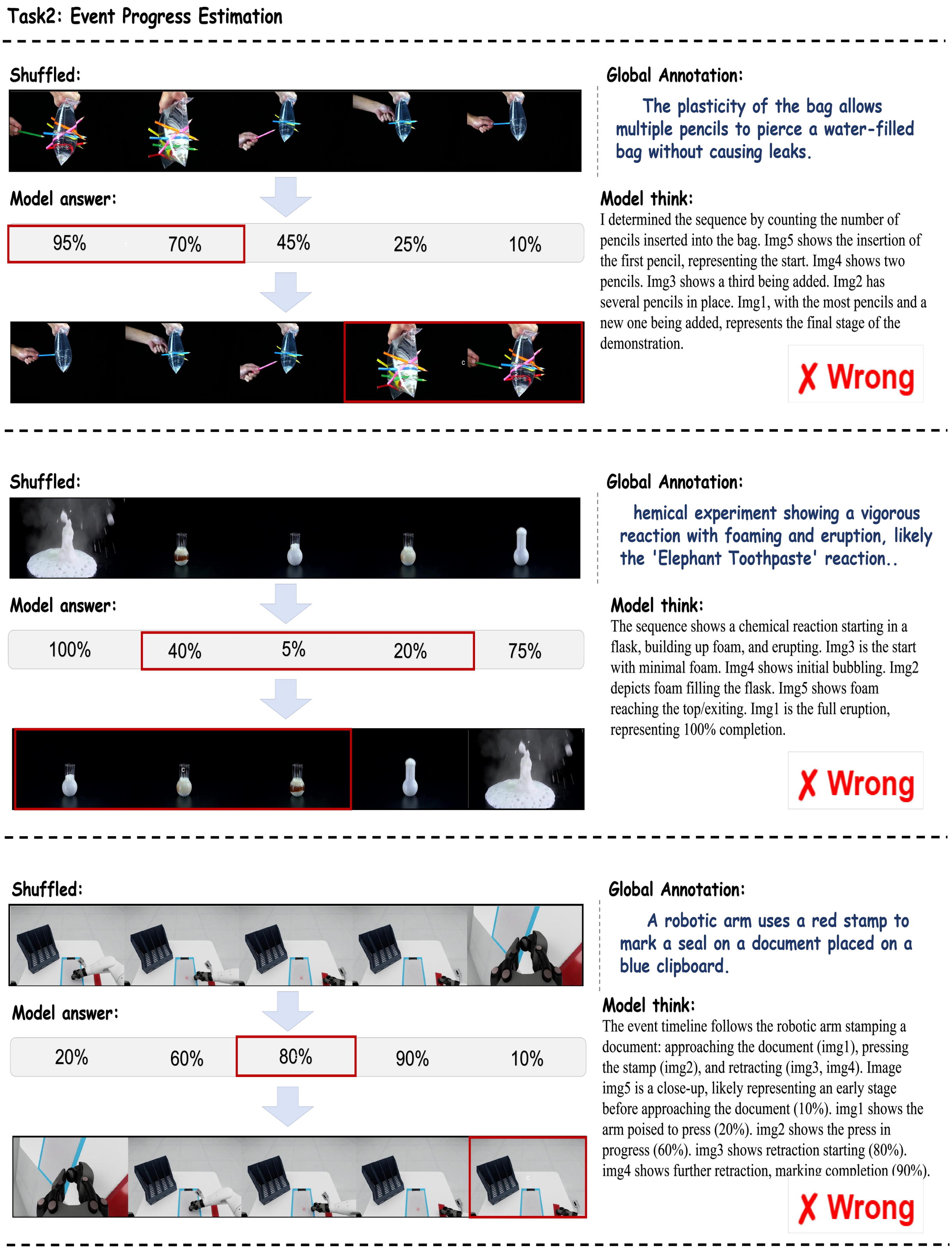

Figure 13: Failure cases in the Event Progress Estimation task.

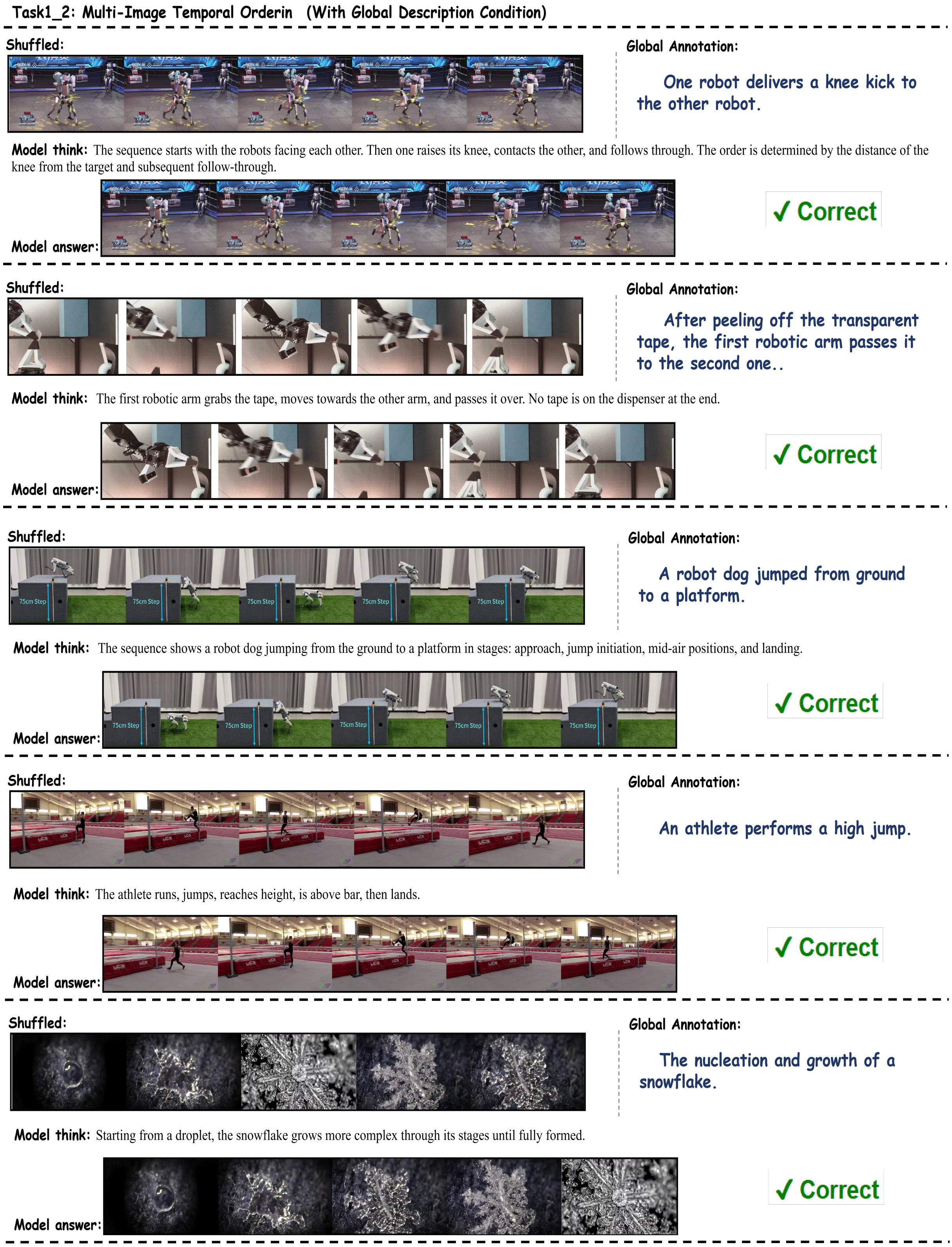

Figure 14: Representative cases from the Multi-Image Temporal Ordering task under the global description condition.

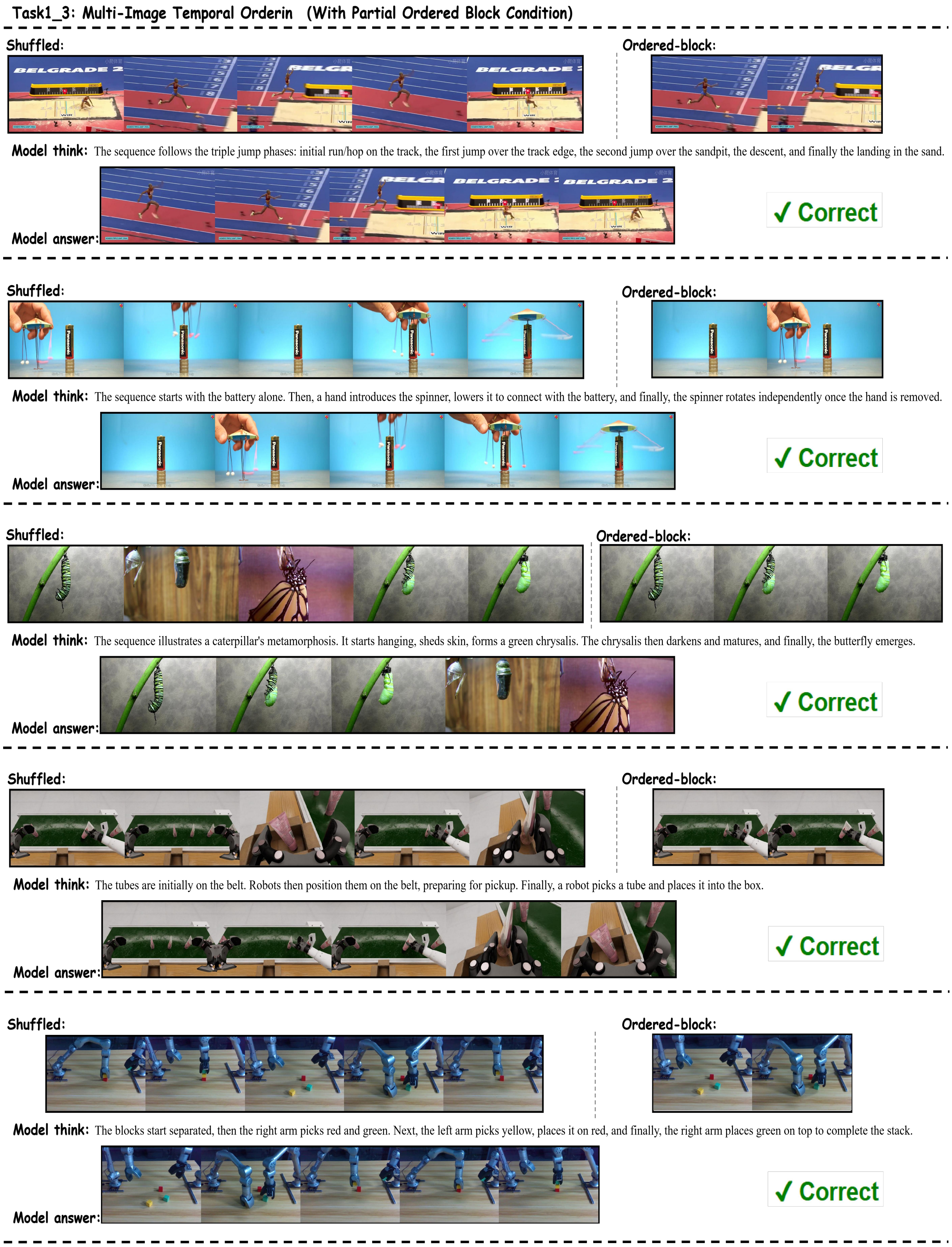

Figure 15: Representative cases from the Multi-Image Temporal Ordering task under the partial ordered block condition.

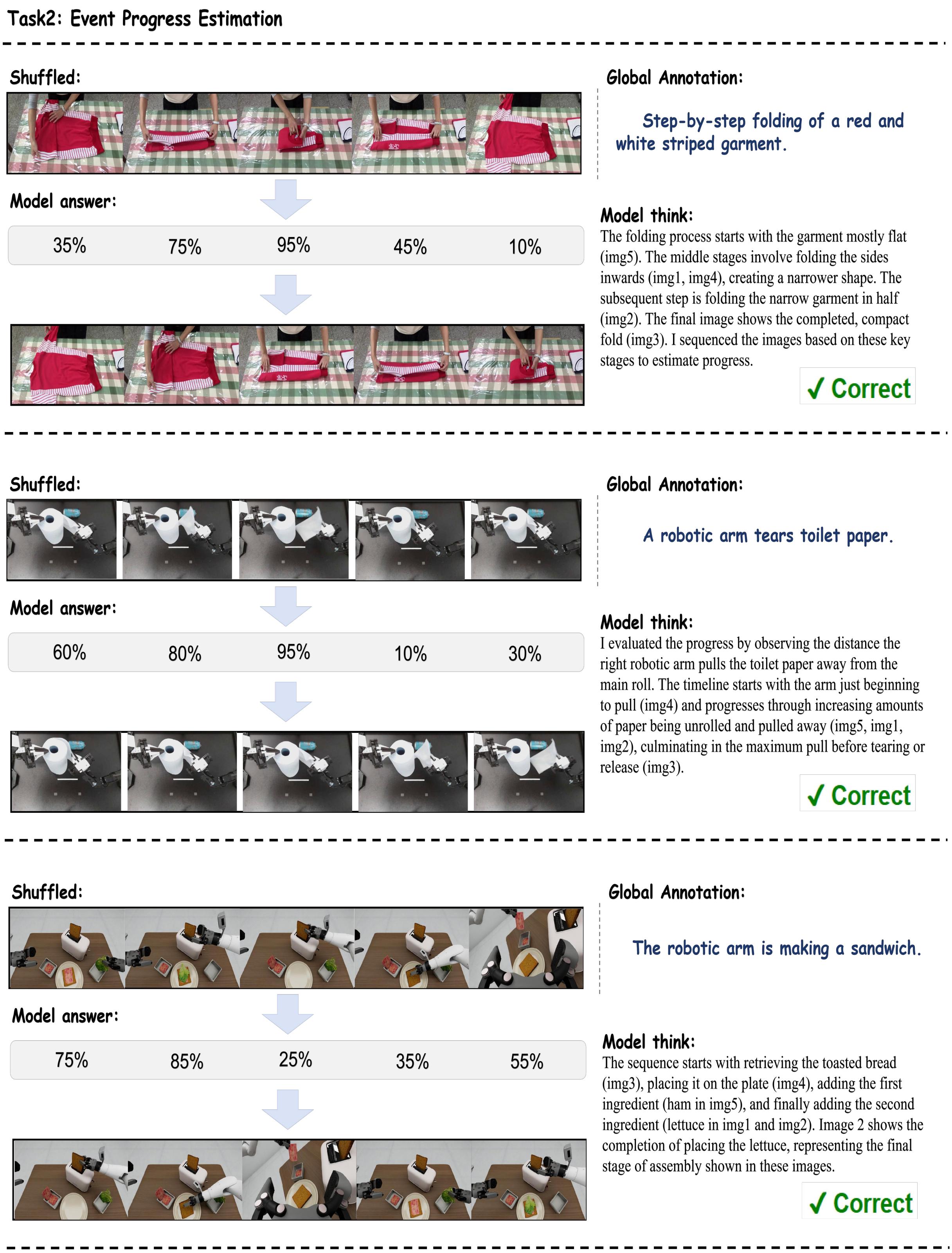

Figure 16: Representative cases from the Event Progress Estimation task.

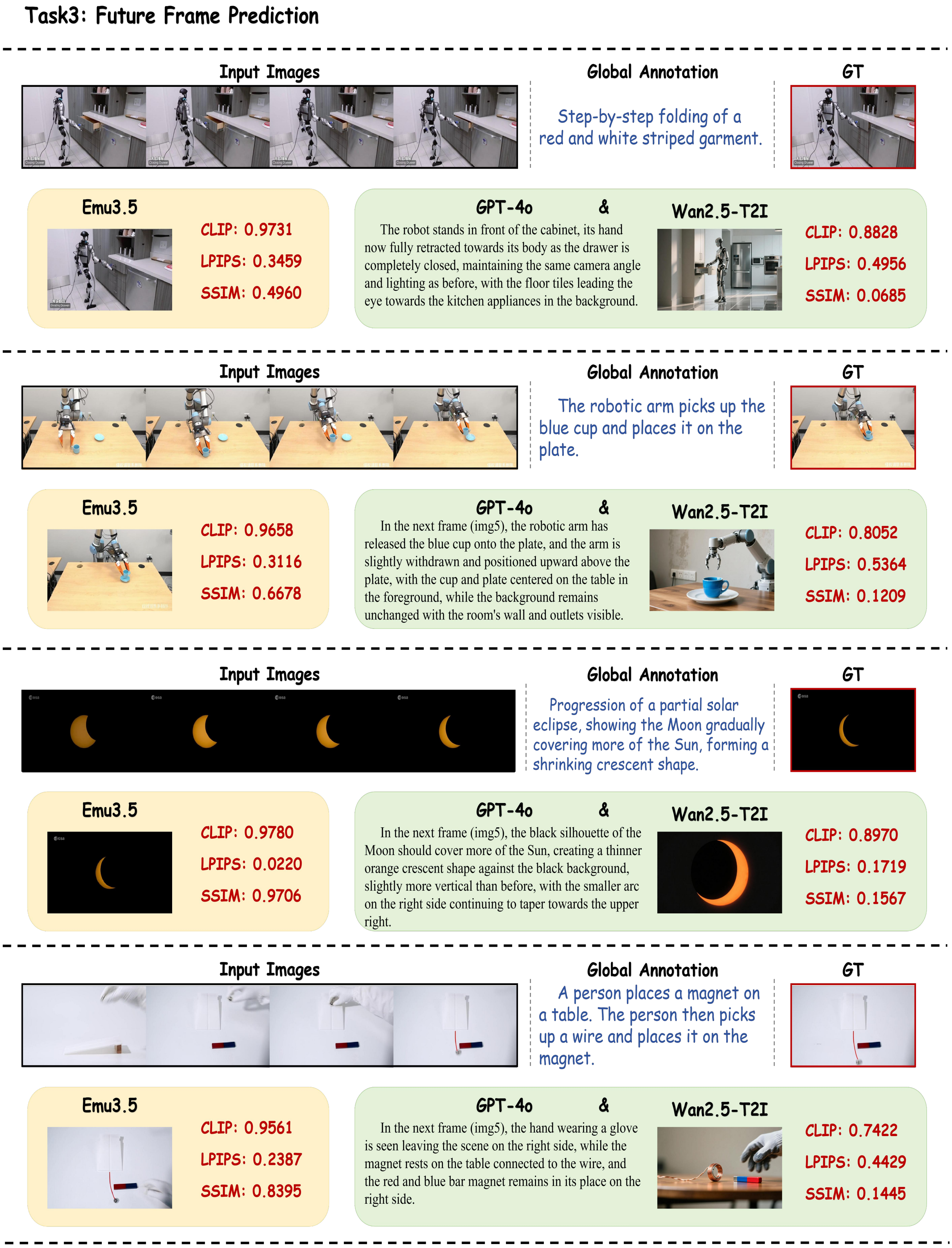

Figure 17: Representative cases from the Future Frame Prediction task.

Multi-Image Temporal Ordering(No-Context Condition) Prompt

```
You are a high-quality visual reasoning assistant.
Task: given five images that portray sequential moments of an event, determine the chronological order (from earliest to latest).
Respond ONLY with a single JSON object, without extra text.

Carefully look at ALL five images provided.
1.The images are given in their original order (index 1 to 5).
2.Do not assume the current order is correct.
3.Decide the proper chronological order of these images, from the earliest moment to the latest.
4.Respond ONLY with a valid JSON object containing:
- "think": a brief explanation of your reasoning (no more than 300 characters).
- "steps": an object with keys "img1" to "img5", where:
   "img1" = the index (1-5) of the image that should be FIRST in chronological order
   "img2" = the index (1-5) of the image that should be SECOND in chronological order
   "img3" = the index (1-5) of the image that should be THIRD in chronological order
   "img4" = the index (1-5) of the image that should be FOURTH in chronological order
   "img5" = the index (1-5) of the image that should be FIFTH in chronological order

Example:
{
   "think": "the cup is empty then filled......",
   "steps": { "img1": 3, "img2": 1, "img3": 2, "img4": 4, "img5": 5}
}
```

Multi-Image Temporal Ordering(With Global Description Condition) Prompt

```
You are a high-quality visual reasoning assistant.
Task: given five images that portray sequential moments of an event, determine the chronological order (from earliest to latest).
Respond ONLY with a single JSON object, without extra text.

Context information (may help you understand the scene):
{context}

Carefully look at ALL five images provided.
1.The images are given in their original order (index 1 to 5).
2.Do not assume the current order is correct.
3.Decide the proper chronological order of these images, from the earliest moment to the latest.
4.Respond ONLY with a valid JSON object containing:
- "think": a brief explanation of your reasoning (no more than 300 characters).
- "steps": an object with keys "img1" to "img5", where:
   "img1" = the index (1-5) of the image that should be FIRST in chronological order
   "img2" = the index (1-5) of the image that should be SECOND in chronological order
   "img3" = the index (1-5) of the image that should be THIRD in chronological order
   "img4" = the index (1-5) of the image that should be FOURTH in chronological order
   "img5" = the index (1-5) of the image that should be FIFTH in chronological order

Example:
{{
   "think": "the cup is empty then filled......",
   "steps": {{ "img1": 3, "img2": 1, "img3": 2, "img4": 4, "img5": 5}}
}}
```

Figure 18: Multi-Image Temporal Ordering(No-Context Condition & With Global Description Condition) Prompt.

Multi-Image Temporal Ordering(With Partial Ordered Block Condition) Prompt

You are a high-quality visual reasoning assistant.
Task: Given five images that portray sequential moments of an event, determine their chronological order (from earliest to latest).
Respond ONLY with a single JSON object, without any extra text.

Instructions:
1. From the section **FRAME BLOCK**, you will be shown a small set of images that are known to be consecutive frames (a continuous block in time). These are provided as visual hints.
2. After that, in the section **SHUFFLE FRAME**, you will see five shuffled images. Your goal is to infer their correct chronological order.
3. Carefully analyze all visual cues and use the continuous block as prior knowledge to reason about the event timeline.
4. Respond ONLY with a valid JSON object containing:
  - "think": a brief explanation of your reasoning (no more than 300 characters)
  - "steps": an object mapping each "img1"–"img5" to the original indices of the five shuffled images, such that:
    "img1" = the index (1–5) of the image that should be FIRST in chronological order
    "img2" = the index (1–5) of the image that should be SECOND
    "img3" = the index (1–5) of the image that should be THIRD
    "img4" = the index (1–5) of the image that should be FOURTH
    "img5" = the index (1–5) of the image that should be FIFTH

Output Example:
```
{
 "think": "The cup is empty first, then filled with water, then placed back on the table.",
 "steps": { "img1": 3, "img2": 1, "img3": 2, "img4": 4, "img5": 5 }
}
```
FRAME BLOCK:

Now observe the following shuffled frames.
SHUFFLE FRAME:

Event Progress Estimation Prompt

You are a high-quality visual reasoning assistant.
Task:
1. You will be given five images depicting sequential moments of a single continuous event.2. Your goal is to estimate how far along the event has progressed in each image.
3. For each image, estimate the completion percentage (0–100%), where:
  - 0% = the event has just started (the very beginning),
  - 100% = the event has fully finished or reached its goal.

Important Notes:
1. The given five images may be shuffled (not in correct chronological order).
2. Each image's percentage should be independent, not necessarily evenly spaced.
3. Carefully compare the five images and judge their relative positions along this timeline.
4. Avoid giving identical percentages unless two images truly represent the same event stage.
5. Use the provided context information to better understand the event nature.

Output Requirements:
Please respond ONLY with a valid JSON object (no additional text or explanations).
Your output must include:
- "think": a concise reasoning summary (≤300 characters) explaining how you estimated the progress.
- "progress": a dictionary mapping each image (img1–img5) to its completion percentage (0–100).

Example Output:
```
{
 "think": "Explain your reasoning briefly.",
 "progress": {
    "img1": 0-100,
    "img2": 0-100,
    "img3": 0-100,
    "img4": 0-100,
    "img5": 0-100
 }}
```

Figure 19: Multi-Image Temporal Ordering(With Partial Ordered Block Condition) Prompt. & Event Progress Estimation Prompt.

Future Frame Prediction(GPT-4o) Prompt

You are a highly capable visual reasoning and image generation assistant.

Task:
You will be given four images (img1–img4) showing sequential moments of a single continuous event.
Your goal is to PREDICT and GENERATE the fifth image (img5) that would logically follow in this sequence.

Guidelines:
1. Carefully observe the visual progression across img1–img4.
2. Understand the event's dynamics, objects, actors, and their interactions.
3. Imagine how the scene will continue in the next moment (img5).
4. Describe not only the actions but also the **visual layout**:
   - Relative positions of objects and actors.
   - Foreground, midground, and background elements.
   - Spatial separation or framing (e.g., left/right/top/bottom placement).
   - Perspective, viewpoint, and camera angle.
   - Scene composition, balance, and any notable visual divisions.
5. Ensure visual consistency with previous frames (colors, lighting, viewpoint, actions).
6. Output ONLY the description that could guide the generation of the next frame (img5). Avoid any extra text or explanation.

Context information:
{context}

Output:
Now, generate a detailed description of the next frame (img5), including **actions, object positions, and overall layout/composition**, in one clear English sentence.

Future Frame Prediction(Emu3.5) Prompt

You are a highly capable visual reasoning and image generation assistant.

Task:
You will be given four images (img1–img4) showing sequential moments of a single continuous event.
Your goal is to PREDICT and GENERATE the fifth image (img5) that would logically follow in this sequence.
Guidelines:
1. Carefully observe the visual progression across img1–img4.
2. Understand the event's dynamics, objects, actors, and their interactions.
3. Imagine how the scene will continue in the next moment (img5).
4. Ensure visual consistency with previous frames (colors, lighting, viewpoint, actions).

Context information:
{context}

Output:
Now, generate the next frame (img5).
(img1~img4)

Figure 20: Future Frame Prediction(GPT-4o) Prompt. & Future Frame Prediction(Emu3.5) Prompt.