EGO-FLIGHT: EGOCENTRIC GROUNDING OF ORDER FOR FRAME-LEVEL INFERENCE IN GENERAL HUMAN TIMELINES

Jiahang He* **Anya Singh** **Jai Relan** **Varun Nair**
Relling Systems

ABSTRACT

Multimodal Large Language Models have shown impressive progress across vision-language tasks, but they still struggle with temporal reasoning, a critical skill for understanding dynamic visual content. We introduce EGO-FLIGHT, a benchmark and dataset designed to directly evaluate temporal reasoning in Vision-Language Models (VLMs) through frame-ordering tasks in human-like, first-person visual contexts. Our dataset contains 1,056 continuous, egocentric video clips that capture natural variations in lighting, motion, and occlusion, providing a first-person perspective that mirrors how humans experience and interpret dynamic scenes. Experiments on frame-sorting tasks varying four controlled variables reveal that current models perform substantially below the human baseline, though longer videos, fewer frames, and more annotation generally improve performance. Finally, applying two LoRA fine-tuning strategies to a VLM trained on our hand-collected data improves performance over the base model, providing a promising path toward enhancing temporal reasoning capabilities. We hope this work advances research on temporal understanding and encourages the development of models that more closely align with human perception while supporting realistic learning in embodied systems such as robots.

1 INTRODUCTION

Multimodal Large Language Models (MLLMs) have recently demonstrated strong performance on a variety of tasks, including visual question answering, image captioning, and scene description (Xue et al., 2024; Caffagni et al., 2024; Kar et al., 2024; Yang et al., 2025a). Despite these successes, significant research is still needed to advance MLLMs toward human-level capabilities (Fu et al., 2024; Wang et al., 2024; Fan et al., 2025). In particular, MLLMs continue to struggle with temporal reasoning, a fundamental skill humans employ when interpreting dynamic visual content (Chen et al., 2024a; Imam et al., 2025; Plizzari et al., 2025). This limitation hinders model performance in understanding temporal information, such as videos, which is abundant and critical in real-world visual data (Cores et al., 2024; Zhou et al., 2025).

This paper aims to address this gap by introducing EGO-FLIGHT: a comprehensive egocentric video dataset and benchmark designed to evaluate the temporal reasoning abilities of Vision-Language Models (VLMs) in human-like, first-person visual contexts. Existing temporal benchmarks face two main challenges. First, many benchmarks assess temporal reasoning indirectly through question answering, which blends content comprehension with reasoning and does not directly measure a model's ability to understand temporal sequences (Li et al., 2023; Swetha et al., 2025). Second, benchmarks that do not rely on QA often use artificially combined or highly curated video clips, which omit the natural variations, noise, and ecological complexity present in real-world perception (Ahn et al., 2025; Ballout et al., 2025). To address these challenges, EGO-FLIGHT evaluates a VLM's ability to temporally sort shuffled frames from an action sequence, supported by a dataset of 1,056 continuous, egocentrically recorded video clips carefully curated by our team. Because these clips are continuous, they capture natural motion, lighting variations, and occlusions that occur in real-world perception. We adopt an egocentric design because its inherent variability and

*Correspondence to jhe.primary@gmail.com

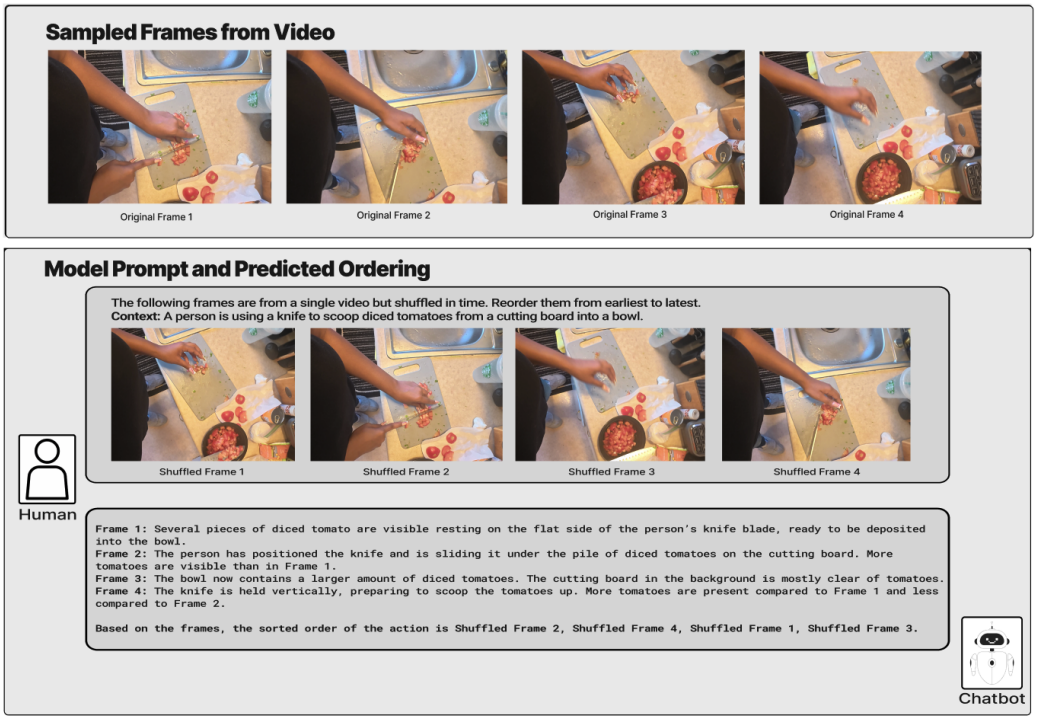

Figure 1: Example interaction from our frame-ordering benchmark. Frames are uniformly sampled from a continuous video depicting an everyday action, then shuffled and presented to a Vision-Language Model together with a prompt requesting chronological reordering and contextual understanding. The model generates frame-level descriptions and predicts the temporal order, which is evaluated against the ground-truth sequence and human baseline performance.

continuous visual flow provides a more human-like model of perception and realism, reflecting the dynamic nature of embodied visual experience. Our benchmark evaluates performance across four experimental variables: number of frames, video length, annotation level, and the amount of hints provided. Results show that current VLMs perform significantly below the human baseline across all nine frame sorting experiments, with a large gap between open-source and closed-source models. We also find that models perform best on longer clips, which provide richer temporal cues, but struggle as the number of frames increases due to subtle or uninformative visual changes between frames. Brief textual annotations substantially improve performance, whereas comprehensive descriptions provide a similar increase for proprietary models but yield little improvement for open-source ones. Finally, we fine-tune Qwen3-VL-8B-Thinking using low-rank adaptation (LoRA) to enhance its temporal reasoning abilities, which we evaluate on a similar frame-sorting task. Two fine-tuning strategies, language-model-only and joint language–vision adaptation, both yield significant gains, with the language-model-only approach achieving the strongest improvements in frame ordering accuracy.

In summary, our paper differs from existing temporal evaluation approaches through the combination of 3 key features. **(1)** We evaluate VLMs on individual frames from a single continuous action, rather than on multi-frame clips spanning multiple actions, allowing our benchmark to isolate temporal reasoning from motion cues or abrupt inter-action transitions. **(2)** We assess each model's ability to sort multiple shuffled frames into chronological order, providing a direct and interpretable measure of complete temporal understanding rather than inferring it indirectly through question answering. **(3)** Our dataset is made up of continuous, egocentric clips that preserve natural variations, such as lighting shifts, occlusions, and camera movement, that humans intuitively process but models often find challenging. By grounding evaluation in egocentric perception, our approach mirrors how humans experience continuous visual input in daily life, linking temporal reasoning to embodied and context-aware understanding. Together, this design, as shown in Figure 1, provides an ecologically valid test of temporal reasoning that minimizes confounding factors, ensuring that model performance reflects genuine temporal understanding. We aim for EGO-FLIGHT to help develop VLMs across

diverse applications, including video understanding, robotics, and real-world reasoning. Links to the code repository and partial dataset are provided in Appendix I.

2 Related Works

VLM Benchmarks. Recent developments in VLMs have led to a growing interest in benchmarking their capabilities across perception, reasoning, and multimodal generation (Bi et al., 2025; Liang et al., 2025; Cheng et al., 2025). Early benchmarks focused on image–text alignment and object grounding, evaluating whether models could correctly associate visual content with descriptive language (Yu et al., 2023; Bai et al., 2023; Chen et al., 2023). Later efforts extended this to compositional reasoning, including understanding spatial relations, attribute binding, and object interactions, revealing limitations in integrating visual features with language (Zeng et al., 2024; Hua et al., 2024; Huang et al., 2024a). More recent benchmarks have emphasized open-ended reasoning and generation, evaluating tasks such as scene description, long-form and complex question answering, and temporal or sequential understanding in short videos (Liu et al., 2024a; Park et al., 2024; Rawal et al., 2024; Zhang et al., 2024; Kong et al., 2025). Collectively, these works reflect a progression from simple recognition and grounding tasks toward more nuanced evaluations that probe reasoning, compositionality, and robustness. Our benchmark builds on this by evaluating temporal reasoning, an essential human cognitive skill, and grounding this evaluation in human perception through continuous, egocentric video sequences that capture natural visual dynamics.

Temporal Understanding and Egocentric Datasets. Temporal reasoning in VLMs has been increasingly scrutinized through benchmarks designed to probe models' understanding of event order, causality, and dynamic changes (Liu et al., 2024b; Wang et al., 2025; Imam et al., 2025; Song et al., 2025). Early efforts revealed that many models rely heavily on static cues rather than genuine temporal understanding, and recent benchmarks collectively examine reasoning across multiple temporal segments, logical relationships between events, and robustness under temporal perturbations (Li et al., 2023; Cores et al., 2024; Shangguan et al., 2024; Feng et al., 2025). VECTOR and SPLICE, in particular, move beyond question-answering by directly evaluating a model's ability to temporally sort shuffled frames, showing that even slight ordering perturbations can significantly alter predictions (Ahn et al., 2025; Ballout et al., 2025). Alongside these developments, egocentric datasets have gained popularity for their human-like perspective and fine-grained capture of everyday actions (Wang et al., 2023; Grauman et al., 2023; Huang et al., 2024b). A few benchmarks have even evaluated VLMs' temporal reasoning specifically using egocentric video data (Mangalam et al., 2023; Yuan et al., 2025; Plizzari et al., 2025). We expand upon these benchmarks by introducing experimental variables, including hint frames and varying frame counts, while demonstrating that fine-tuning can enhance model performance. Additionally, although our benchmark shares VECTOR and SPLICE's focus on frame-ordering evaluation, it differs by using continuous, egocentric videos rather than curated clips, and samples discrete frames across full sequences to isolate reasoning from motion cues.

3 Benchmark Design

3.1 Data Collection

Our dataset is curated from egocentric videos recorded by our team, depicting people performing everyday tasks and capturing the temporal dynamics of real-world actions. Adopting an egocentric perspective provides continuous, naturally varying footage with realistic occlusions and viewpoint shifts, making temporal reasoning more challenging while closely mirroring human visual experience. For example, the top sequence of Figure 2 shows a person transferring food from a skillet to a metal plate: as the plate gradually fills and the skillet empties, models must track temporal progression and shifting viewpoints to correctly interpret the action. Other examples, such as restocking a vending machine, involve multi-step actions that demand models to track changes over time through evolving visual context. In contrast to prior benchmarks that group multiple frames from different actions into short clips and evaluate models on ordering those clips, our setup requires models to order individual frames within a single continuous action from a single clip, isolating temporal reasoning from inter-action transitions. This design removes explicit motion continuity between clips, making

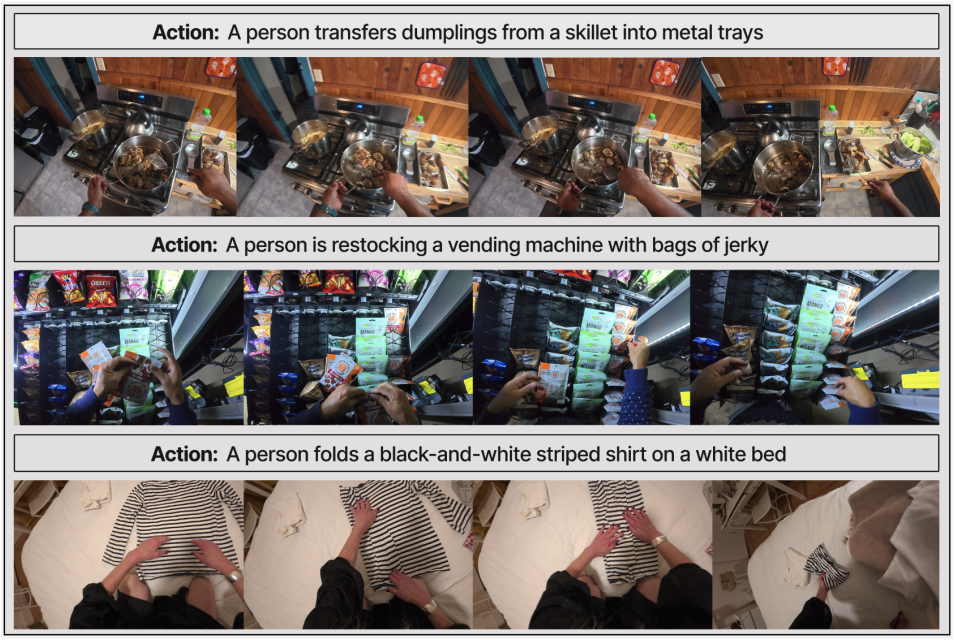

Figure 2: Example clips and their sampled frames from our dataset. Each sequence illustrates a distinct temporal progression: (1) food being transferred from a skillet, (2) jerky being restocked in a vending machine, and (3) clothes being folded. The visual context shown here differs from the textual annotation provided to the model, which is presented in Appendix F

.

the task more challenging since models must infer temporal order from subtle visual changes rather than relying on clear transitions or scene-level cues. Success on our benchmark requires models to explicitly infer the chronological order between frames, even when visual changes are subtle or occluded, closely approximating how humans perceive events in the real world.

Each video in our dataset was recorded by humans using cap-mounted egocentric cameras in 4K at 60 FPS. Participants came from diverse backgrounds, including cooks, farmers, and convenience store workers, resulting in a total of approximately 240 hours of raw video. To curate clips suitable for evaluating temporal reasoning, human annotators reviewed the raw recordings and selected segments containing meaningful temporal cues, such as changes in the primary action or evolving activity in the background. While the egocentric perspective naturally introduces shifts in camera angle, we ensured that these variations did not substantially alter the scene, preserving temporal continuity for the task. As summarized in Appendix H, we quantitatively characterize our dataset to illustrate its diversity, deliberately avoiding the overrepresentation of simple or repetitive tasks to capture a broad spectrum of interactions and activities. Figure 2 further illustrates this diversity, showcasing examples such as folding clothes, restocking a vending machine, and other everyday actions. The dataset includes both indoor and outdoor settings, with interactions involving single and multiple people, enabling evaluation across a broad range of human activities. While initial annotations were generated using Gemini 3.0 Flash, all annotations were closely reviewed and revised by human annotators to ensure accuracy and prevent model-specific bias. Each clip includes two types of annotations: a concise label-based annotation and a more comprehensive, detailed annotation capturing temporal progression. In total, our dataset consists of 1,056 clips, which were selected from 240 hours of raw footage spanning 195 distinct action tasks. Clip durations range from 5 to 85 seconds and include detailed annotations capturing scene actions, providing a rich and diverse resource for evaluating temporal reasoning in vision-language models.

3.2 Experimental Variables

Appendix E details the nine experiments we conducted, covering all variations of the experimental variables. Each variable is systematically varied as follows:

- Number of Frames: We test each VLM on 2 frames, 4 frames, and 8 frames. While increasing the number of frames provides the model with greater temporal context, it makes the sorting task significantly more challenging because a larger set of frames contains a higher proportion of uninformative or redundant frames.
- Clip Length: We categorize clips into three durations: short (5–10 seconds), medium (10–25 seconds), and long (over 25 seconds). While longer clips cover a greater temporal span, which can make ordering more difficult, the extended transitions also provide richer temporal cues that can help the model determine the correct frame sequence.
- Annotation descriptiveness: We vary the level of annotation given to the model across three conditions: no annotation, brief annotation, and comprehensive annotation. Brief annotation is seen in Appendix F.1, the model is given a single-line summary of the task being performed. Appendix F.2 shows examples of comprehensive annotation, the model is given a detailed step-by-step description of the actions in the clip. Prior work has shown that brief annotations can improve model performance on tasks involving temporal reasoning; here, we investigate whether providing comprehensive annotations produces a similar improvement in frame-sorting tasks (Cai et al., 2024; Chen et al., 2024b).
- Hints: We provide models with two reference frames from the original sequence, reducing the task to sorting the remaining frames relative to these hints. For 4-frame sequences, we systematically test two hint configurations, providing either the first and third frames or the second and fourth frames. This effectively transforms the problem into a 2-frame sorting task with additional contextual references.

The baseline configuration corresponds to experiments using four frames sampled from medium-length videos, with brief annotations and no hints provided. In the data table, we refer to this baseline configuration as the original experiment.

3.3 Models Used

We evaluate a combination of proprietary and open-source VLMs, encompassing a variety of sizes and architectures to provide a comprehensive overview of model performance. In total, we use 9 models including GPT-4o, Claude 4.5 Sonnet, Gemini 3.0 Flash, Qwen3-VL-8B-Thinking, Qwen3-VL-32B-Thinking, LLaVA-v1.6-Vicuna-13B, Phi 3.5 Vision Instruct, R-4B, and Gemma 3 4B (Hurst et al., 2024; Anthropic, 2025; Google, 2025; Yang et al., 2025a; Liu et al., 2023; Abdin et al., 2024; Yang et al., 2025b; Kamath et al., 2025). All models are tested with a temperature of 0 for deterministic answers.

3.4 Testing Pipeline and Model Task

The general outline of our experiment is to provide the model with frames from a single action shuffled in temporal order, and prompt it to reconstruct the original sequence. Frames are extracted from each clip sequentially and sampled evenly to balance temporal coverage with model input length, ensuring that key transitional moments are captured without overloading the model. They are then downsampled to 224×224 pixels and fed into the model as individual images in a sequence. Each set of frames is combined with text containing task instructions, scene context, and output format requirements into a single prompt, shown in Appendix G. Findings from our permutation robustness analysis in Appendix A show that models are highly sensitive to input frame order and that performance varies substantially across different frame permutations. To account for this effect, we randomly shuffled the indices of frames for each clip to generate the initial input order to the model. We then recorded the specific shuffled orders used for each question when evaluating the first model and reused these exact orders for all subsequent models. Randomizing input permutations ensures that performance variations across permutations average out, while using the same randomized inputs across models ensures that differences in performance reflect genuine model variation rather than inconsistencies in input ordering. Each experiment is evaluated on 156 clips per model to provide broad coverage of the dataset's diversity.

As shown in Figure 1 our evaluation pipeline proceeds as follows: (1) feed the evenly sampled and shuffled frames and accompanying text to the model, (2) capture the model's output, which may include predicted action labels, temporal descriptions, and reasoning traces, (3) compare predictions

against ground-truth order, and (4) compute evaluation metrics such as accuracy and reasoning consistency while also comparing performance to a human baseline. The human baseline consists of 15 trained annotators, each assigned 102 questions spanning two of the nine experimental conditions, resulting in a total of 1,530 baseline evaluations. To collect these annotations, we developed a web-based interface that allowed participants to independently complete their assigned experiments. Each annotator received detailed written instructions and several practice examples to ensure a clear understanding of the task before beginning. The platform automatically prevented participants from completing overlapping versions of the same clip, such as both the original 4-frame and 4-frame-with-hints conditions. Upon submission, response times were recorded for every question, and submissions from annotators who spent unreasonably little time or showed clear misunderstanding of the task were excluded from analysis.

3.5 Evaluation Metrics

To evaluate model performance on frame sorting tasks, we use standard evaluation metrics including: Pairwise Accuracy, Mean Absolute Distance (MAD), Longest Common Subsequence Ratio (LCS Ratio), Edit Distance, and Exact Match. Results for these metrics are shown in Appendix D.

More uniquely, we evaluate performance using Kendall's Tau. Kendall's Tau directly measures the degree of pairwise agreement between a model's predicted order of events and the true chronological order. For a sequence of n elements, it considers all $\frac{n(n-1)}{2}$ possible pairs and determines how many are correctly ordered relative to the ground truth. We use Kendall's Tau-b specifically, but since all frames have distinct positions, the variant choice does not affect results.

Formally, Kendall's Tau-b is defined as:

$$\tau = \frac{(C - D)}{C + D},$$

where C and D denote the number of concordant and discordant pairs, respectively. The value of τ ranges from -1 to $+1$, offering intuitive interpretability:

- A positive value (e.g., $\tau = 0.8$) indicates strong agreement, meaning most pairwise orderings are correct.
- A value near 0 reflects random ordering—effectively the random baseline, since random guesses yield roughly equal numbers of concordant and discordant pairs.
- A negative value (e.g., $\tau = -0.5$) indicates systematic inversion, where a model tends to predict the opposite of the true order.

When evaluating frame-ordering tasks, metrics like Exact Match are highly sensitive to sequence length, since perfectly sorting all frames becomes exponentially harder as the number of frames increases. Kendall's Tau, in contrast, measures the relative ordering of all pairs of frames rather than requiring a full sequence match. This means that even if a model makes some mistakes, it can still receive partial credit for correctly ordering most frame pairs allowing for a more nuanced assessment of how models handle temporal relationships. Additionally, the fact that the random baseline for Kendall's Tau is 0.0 provides an intuitive reference point for comparison as seen in Table 1.

4 Results and Analysis

Performance across experimental variables revealed consistent trends in how models handled temporal reasoning. As shown in Table 1, accuracy generally decreased as the number of frames increased, with GPT-4o specifically dropping from 0.29 on 2-frame sequences to 0.23 on 4-frame sequences and 0.16 on 8-frame sequences. This aligns with the human baseline and reflects how adding more frames introduces subtler or more redundant transitions, making it increasingly difficult to infer temporal relationships. However, some smaller open-source models, such as Phi 3.5 and R-4B displayed near-random performance even on two-frame sequences, indicating limited temporal understanding. Video length also played an important role: although longer clips introduced greater variation in camera angle and lighting, they provided richer temporal cues and were generally easier for models to interpret than shorter and medium ones. The human baseline also performed slightly better on

Table 1: Performance of evaluated models across frame-sorting experiments, reported using the Kendall's Tau-b metric. The original experiment corresponds to the 4-frame, medium-length video condition with brief annotations and no hints.

Experiment	GPT-4o	Claude-Sonnet-4.5	Gemini-3.0-Flash	Gemma R-4B 4B		Phi 3.5 Vision	Llava-13B	Qwen3-VL-8B	Qwen3-VL-32B	Human Baseline
Original	0.23	0.09	**0.35**	-0.01	0.04	-0.01	0.07	0.14	0.14	0.54
Changing Frames										
2 Frames	0.29	0.33	**0.35**	0.29	0.04	-0.09	0.01	0.08	0.18	0.74
8 Frames	0.16	0.09	**0.29**	0.03	0.00	-0.01	0.05	0.07	0.16	0.37
Changing Video Length										
Short	0.16	0.11	**0.26**	-0.03	0.05	-0.03	0.04	0.09	0.11	0.41
Long	0.26	0.12	**0.31**	0.12	-0.04	0.00	-0.03	0.16	0.17	0.43
Changing Annotation Description										
None	0.20	0.05	**0.30**	0.02	0.01	0.00	0.00	0.06	0.18	0.48
Comprehensive	0.27	0.15	**0.44**	0.01	0.03	0.07	0.08	0.11	0.12	0.55
Changing Hints										
Frame 1 & 3	0.28	0.50	**0.65**	0.12	-0.01	-0.01	-0.08	0.33	0.46	0.79
Frame 2 & 4	0.45	0.31	**0.61**	-0.02	-0.02	0.09	0.10	0.41	0.48	0.79

long clips compared to short clips, 0.43 compared to 0.41, yet achieved its highest score of 0.54 on medium-length ones. These results suggest that while models benefit from abundant temporal cues, humans perform best when sequences provide a balance between temporal span and informative content.

Annotation quality also had a significant impact on model performance. The introduction of brief annotations led to a substantial improvements across virtually all models, highlighting the importance of textual guidance in aligning model predictions with video context. While the human baseline also improved with brief annotations, the gains were smaller, reflecting humans' ability to infer actions directly from visual and contextual cues without heavy reliance on textual descriptions. Interestingly, performance also improved for proprietary models when moving from brief to comprehensive annotations, with Gemini 3.0 Flash increasing from 0.35 to 0.44, suggesting that these models can effectively interpret and leverage more descriptive prompts. On the other hand, open-source models showed little improvement, and even slight degradation, indicating that excessively long or comprehensive textual inputs may overwhelm smaller models rather than enhance their understanding. For the human baseline, performance also remained stable at around 0.55, indicating that comprehensive annotations are generally unnecessary for humans to understand everyday action sequences.

In hint experiments, two frames are pre-ordered, so metric evaluation was performed only on the two unsorted frames. Providing hint frames proved beneficial across models, though the extent of improvement varied. Models with lower initial performance, such as Qwen 32B, exhibited the most significant gains, nearly tripling their two-frame accuracy from 0.18 to 0.48 when hint frames were provided. Stronger proprietary models also benefited, though to a lesser degree. Interestingly, while the human baseline score in hint-based experiments remained comparable to the two-frame condition, models exhibited substantial improvement. This may partly reflect their lower starting accuracy, but also suggests that models effectively leverage the reference frames to anchor their predictions, demonstrating genuine temporal reasoning. The relative benefit of providing hints in positions 1 and 3 versus 2 and 4 varied, indicating that no single hint configuration is superior; effectiveness depends on the model's existing temporal understanding and internal representation of the sequence.

Overall, proprietary models consistently outperformed open-source ones across all experiments, with Gemini 3.0 Flash achieving a score of 0.35 on the original 4-frame experiment, whereas the best-performing open-source model Qwen 32B reached only 0.14. Among the commercial systems, Gemini 3.0 Flash achieved the strongest results in all categories, while GPT-4o performed slightly better than Claude-Sonnet 4.5. In contrast, many smaller open-source models struggled with nearly

all tasks, highlighting the limitations of smaller architectures in handling complex temporal reasoning. The Qwen models stood out among open-source systems: both Qwen 8B and Qwen 32B demonstrated competitive performance, with the 32B model even surpassing GPT-4o and Claude-Sonnet-4.5 in some of the hint-based experiments. Despite these results, all models still lagged behind human performance. Although Gemini occasionally approached human-level accuracy in a few categories, such as 8 frame experiments, it fell notably short in others, demonstrating that current vision-language models, while capable, still lack robust generalization in fine-grained temporal reasoning.

5 FINE-TUNING

We explore two low-rank adaptation (LoRA) fine-tuning strategies to enhance temporal reasoning in vision-language models using Qwen3-VL-8B-Thinking as the base-model architecture. This model was selected because it is open-source, demonstrated above-random performance on our benchmark, and thus provides meaningful room for improvement.

5.1 FINE-TUNING APPROACHES

LM-Only LoRA Fine-Tuning. In this setup, only the language model component of Qwen3-VL-8B-Thinking is fine-tuned, while the vision encoder remains completely frozen. LoRA adapters are applied to the `q_proj` and `v_proj` layers, corresponding to the query and value projection layers in the attention blocks. Two frames are provided to the model along with a textual prompt asking, "Does Frame A or Frame B happen first?" The images are processed by the frozen vision encoder into embeddings, which are concatenated with text embeddings and passed to the LLM. The model is trained to output either "A" or "B" as a single token. During training, we extract the logits corresponding to the "A" token and "B" token at the final generation position, stack them into a two-class tensor, and compute cross-entropy loss against the ground truth label (1 if A precedes B, 0 otherwise). This approach enables the language model to learn temporal reasoning patterns given static visual embeddings. The LLM learns how to interpret visual feature representations in combination with textual cues and scene context to infer which frame occurs earlier in time.

Joint LM + Vision Encoder LoRA Fine-Tuning. The second approach fine-tunes both the language model and vision encoder simultaneously. As shown in Figure 4, LoRA adapters are applied to the `q_proj` and `v_proj` layers of the LLM, as well as to the `qkv`, `proj`, `fc1`, and `fc2` layers in the vision encoder. The input format and loss function are identical to the LM-Only setup. By updating both components, this approach allows the vision encoder to learn feature representations that are directly optimized for temporal ordering, while the LLM simultaneously learns to interpret these evolving features. This joint fine-tuning fosters co-adaptation between visual and language representations for improved temporal reasoning.

Common Setup. All experiments use the Qwen3-VL-8B-Thinking base model with 4-bit NF4 quantization to fit within GPU memory constraints. To perform fine-tuning, we created a separate egocentric dataset comprising approximately 4,800 briefly annotated frame pairs, entirely disjoint from the videos used in our benchmark. Each 4-frame clip produces six unique frame pairs, and by considering both possible temporal directions, each yields twelve training examples. Of these 4,800 pairs, 1,000 were held out for validation and testing, leaving 3,800 pairs for training. Both fine-tuning strategies were trained for 12 epochs on a single NVIDIA A100-SXM4 GPU.

5.2 EVALUATION METHODOLOGY

To evaluate the temporal reasoning capabilities of our fine-tuned models, we adopt a slightly modified version of our benchmark that leverages pairwise frame comparisons. Since the models were trained with a pairwise temporal objective, asking them to sort all frames at once can be challenging and prone to compounding errors. Breaking the task into binary comparisons simplifies evaluation, mitigates positional biases by randomizing frame presentation as "A" or "B," and allows us to leverage the model's confidence via logits rather than generated text.

For each clip, all $\binom{N}{2}$ frame pairs are generated. Each pair is randomized in presentation, accompanied by the scene context from metadata, and the model is prompted to select the frame that occurred earlier. We adopt this pairwise approach because the model was fine-tuned specifically to predict,

Table 2: Performance of base and fine-tuned Qwen3-VL-8B-Thinking models across frame-sorting experiments with varying numbers of frames, measured with Kendall's Tau rank correlation.

Frame Count	Base Model	LM-Tuned	Joint-Tuned
2 Frames	0.08	**0.23**	0.12
4 Frames	0.03	**0.17**	0.13
8 Frames	0.03	**0.11**	0.08

for any given pair, which frame comes first based on the higher logit for "A" or "B." The logits corresponding to "A" and "B" are compared, with the higher logit indicating the predicted earlier frame. If the frames were swapped for randomization, the prediction is mapped back to the correct original frame indices. Each frame accumulates a "win count," representing the number of pairwise comparisons in which it was predicted as occurring first. After all pairwise comparisons are processed, frames are ranked by their win counts, with ties resolved using the head-to-head result of the tied frames. This procedure naturally scales to clips with more frames; for example, 8-frame clips require 28 pairwise comparisons, with maximum wins for the earliest frame and zero for the latest. For the experiments in this section, each model is evaluated on 100 clips per experiment. The predicted sequences are evaluated against ground truth using Exact Match and Kendall's Tau, with Kendall's Tau serving as the primary measure of temporal ordering accuracy. Together, these two metrics capture both coarse and fine-grained aspects of temporal reasoning performance.

5.3 Results

We evaluated both fine-tuning strategies on the modified temporal ordering experiments described earlier. As shown in Table 2, both approaches improved performance over the pretrained Qwen3-VL-8B-Thinking baseline, demonstrating that targeted adaptation can enhance temporal reasoning in VLMs. The LM-Only LoRA fine-tuning yielded the strongest gains, with Kendall's Tau increasing from 0.03 to 0.17 for 4-frame clips, indicating that even when the vision encoder remains frozen, the language model can learn to interpret static visual embeddings in a temporally meaningful way. By optimizing the query and value projection layers, the LM becomes more adept at reasoning about visual relationships expressed through embeddings, effectively learning temporal patterns in contextual cues.

The Joint LM + Vision Encoder LoRA strategy also improved performance relative to the baseline, achieving a Kendall's Tau of 0.13 on 4-frame clips, providing evidence that co-adaptation between the language and vision components can enhance temporal reasoning. However, its gains were smaller than the LM-Only approach, likely due to the limited size of our dataset, which constrains the model's ability to safely fine-tune the vision encoder without overfitting (see limitations section in Appendix B). Another way to observe improvements for 4-frame clips is through the Exact Match metric: the base model correctly predicted only 3 sequences, whereas both LM-Only and Joint fine-tuned models predicted 11 exact sequences correctly.

Taken together, these results demonstrate that VLMs are capable of learning temporal reasoning behaviors through fine-tuning. The fact that improvements arise even from lightweight LoRA adaptation highlights that temporal understanding can be developed efficiently, without retraining the entire model from scratch. Moreover, the differing outcomes between LM-only and joint fine-tuning emphasize the importance of dataset size when adapting vision encoders, providing practical guidance for future work.

6 Conclusion

In this work, we presented EGO-FLIGHT, a comprehensive egocentric video dataset and benchmark designed to evaluate temporal reasoning in vision-language models. By focusing on the chronological ordering of individual frames from continuous, first-person recordings, our benchmark isolates temporal reasoning from motion cues and provides a direct, interpretable measure of model performance. Our experiments show that existing VLMs struggle with this task, particularly as the number of frames increases or visual changes become subtle, highlighting a gap between current models and

human-level performance. We further demonstrate that two LoRA fine-tuning strategies applied on a separately curated training dataset improves temporal reasoning abilities. These results underscore the importance of high-quality, task-specific datasets for enhancing temporal understanding and we hope our work advances future research in real-world reasoning tasks that more closely align VLMs with human temporal comprehension.

REFERENCES

Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Hassan Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Singh Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Caio C'esar Teodoro Mendes, Weizhu Chen, Vishrav Chaudhary, Parul Chopra, Allison Del Giorno, Gustavo de Rosa, Matthew Dixon, Ronen Eldan, Victor Fragoso, Dan Iter, Abhishek Goswami, Suriya Gunasekar, Emman Haider, Junheng Hao, Russell J. Hewett, Jamie Huynh, Mojan Javaheripi, Xin Jin, Piero Kauffmann, Nikos Karampatziakis, Dongwoo Kim, Young Jin Kim, Mahoud Khademi, Lev Kurilenko, James R. Lee, Yin Tat Lee, Yuanzhi Li, Chen Liang, Weishung Liu, Eric Lin, Zeqi Lin, Piyush Madan, Arindam Mitra, Hardik Modi, Anh Nguyen, Brandon Norick, Barun Patra, Daniel Perez-Becker, Thomas Portet, Reid Pryzant, Heyang Qin, Marko Radmilac, Liliang Ren, Corby Rosset, Sambudha Roy, Olli Saarikivi, Amin Saied, Adil Salim, Michael Santacroce, Shital Shah, Ning Shang, Hiteshi Sharma, Xianmin Song, Olatunji Ruwase, Praneetha Vaddamanu, Xin Wang, Rachel Ward, Guanhua Wang, Philipp Andre Witte, Michael Wyatt, Can Xu, Jiahang Xu, Sonali Yadav, Fan Yang, Ziyi Yang, Donghan Yu, Cheng yuan Zhang, Cyril Zhang, Jianwen Zhang, Li Lyna Zhang, Yi Zhang, Yunan Zhang, Xiren Zhou, and Yifan Yang. Phi-3 technical report: A highly capable language model locally on your phone. *ArXiv*, abs/2404.14219, 2024. URL https://api.semanticscholar.org/CorpusID:269293048.

Daechul Ahn, Yura Choi, Hyeonbeom Choi, Seongwon Cho, San Kim, and Jonghyun Choi. What happens when: Learning temporal orders of events in videos. 2025. URL https://api.semanticscholar.org/CorpusID:283722234.

Anthropic. Introducing claude sonnet 4.5. https://www.anthropic.com/news/claude-sonnet-4-5, 2025. Accessed: 2026-01-13.

Shuai Bai, Shusheng Yang, Jinze Bai, Peng Wang, Xing Zhang, Junyang Lin, Xinggang Wang, Chang Zhou, and Jingren Zhou. Touchstone: Evaluating vision-language models by language models. *ArXiv*, abs/2308.16890, 2023. URL https://api.semanticscholar.org/CorpusID:261397179.

Mohamad Ballout, Okajevo Wilfred, Seyedalireza Yaghoubi, Noha Abdelmoneim, Julius Mayer, and Elia Bruni. Can you splice it together? a human curated benchmark for probing visual reasoning in vlms. *ArXiv*, abs/2509.24640, 2025. URL https://api.semanticscholar.org/CorpusID:281675779.

Jing Bi, Junjia Guo, Susan Liang, Guangyu Sun, Luchuan Song, Yunlong Tang, Jinxi He, Jiarui Wu, Ali Vosoughi, Chen Chen, and Chenliang Xu. Verify: A benchmark of visual explanation and reasoning for investigating multimodal reasoning fidelity. *ArXiv*, abs/2503.11557, 2025. URL https://api.semanticscholar.org/CorpusID:277043352.

Davide Caffagni, Federico Cocchi, Luca Barsellotti, Nicholas Moratelli, Sara Sarto, Lorenzo Baraldi, Marcella Cornia, and Rita Cucchiara. The revolution of multimodal large language models: A survey. In *Annual Meeting of the Association for Computational Linguistics*, 2024. URL https://api.semanticscholar.org/CorpusID:267759688.

Mu Cai, Reuben Tan, Jianrui Zhang, Bocheng Zou, Kai Zhang, Feng Yao, Fangrui Zhu, Jing Gu, Yiwu Zhong, Yuzhang Shang, Yao Dou, Jaden Park, Jianfeng Gao, Yong Jae Lee, and Jianwei Yang. Temporalbench: Benchmarking fine-grained temporal understanding for multimodal video models. *ArXiv*, abs/2410.10818, 2024. URL https://api.semanticscholar.org/CorpusID:273346169.

Jr-Jen Chen, Yu-Chien Liao, Hsi-Che Lin, Yu-Chu Yu, Yen-Chun Chen, and Yu-Chiang Frank Wang. Rextime: A benchmark suite for reasoning-across-time in videos. *ArXiv*, abs/2406.19392, 2024a. URL https://api.semanticscholar.org/CorpusID:270764948.

Zhihong Chen, Ruifei Zhang, Yibing Song, Xiang Wan, and Guanbin Li. Advancing visual grounding with scene knowledge: Benchmark and method. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15039–15049, 2023. URL https://api.semanticscholar.org/CorpusID:260091751.

Zhuo Chen, Zhao Zhang, Zixuan Li, Fei Wang, Yutao Zeng, Xiaolong Jin, and Yongjun Xu. Self-improvement programming for temporal knowledge graph question answering. In *International Conference on Language Resources and Evaluation*, 2024b. URL https://api.semanticscholar.org/CorpusID:268856632.

Kanzhi Cheng, Wenpo Song, Jiaxin Fan, Zheng Ma, Qiushi Sun, Fangzhi Xu, Chenyang Yan, Nuo Chen, Jianbing Zhang, and Jiajun Chen. Caparena: Benchmarking and analyzing detailed image captioning in the llm era. *ArXiv*, abs/2503.12329, 2025. URL https://api.semanticscholar.org/CorpusID:277066648.

Daniel Cores, Michael Dorkenwald, Manuel Mucientes, Cees G. M. Snoek, and Yuki M. Asano. Lost in time: A new temporal benchmark for videollms. 2024. URL https://api.semanticscholar.org/CorpusID:273233203.

Zhiwen Fan, Jian Zhang, Renjie Li, Junge Zhang, Runjin Chen, Hezhen Hu, Kevin Wang, Huaizhi Qu, Dilin Wang, Zhicheng Yan, Hongyu Xu, Justin Theiss, Tianlong Chen, Jiachen Li, Zhengzhong Tu, Zhangyang Wang, and Rakesh Ranjan. Vlm-3r: Vision-language models augmented with instruction-aligned 3d reconstruction. *ArXiv*, abs/2505.20279, 2025. URL https://api.semanticscholar.org/CorpusID:279449925.

Bo Feng, Zhengfeng Lai, Shiyu Li, Zizhen Wang, Simon Wang, Ping Huang, and Meng Cao. Breaking down video llm benchmarks: Knowledge, spatial perception, or true temporal understanding? *ArXiv*, abs/2505.14321, 2025. URL https://api.semanticscholar.org/CorpusID:278769206.

Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A. Smith, Wei-Chiu Ma, and Ranjay Krishna. Blink: Multimodal large language models can see but not perceive. *ArXiv*, abs/2404.12390, 2024. URL https://api.semanticscholar.org/CorpusID:269214091.

Google. Introducing gemini 3: A new era of intelligence. https://blog.google/products-and-platforms/products/gemini/gemini-3/, 2025. Accessed: 2026-01-13.

Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, Eugene Byrne, Zachary Chavis, Joya Chen, Feng Cheng, Fu-Jen Chu, Sean Crane, Avijit Dasgupta, Jing Dong, María Escobar, Cristhian Forigua, Abrham Kahsay Gebreselasie, Sanjay Haresh, Jing Huang, Md Mohaiminul Islam, Suyog Dutt Jain, Rawal Khirodkar, Devansh Kukreja, Kevin J. Liang, Jia-Wei Liu, Sagnik Majumder, Yongsen Mao, Miguel Martin, Effrosyni Mavroudi, Tushar Nagarajan, Francesco Ragusa, Santhosh K. Ramakrishnan, Luigi Seminara, Arjun Somayazulu, Yale Song, Shan Su, Zihui Xue, Edward Zhang, Jinxu Zhang, Angela Castillo, Changan Chen, Xinzhu Fu, Ryosuke Furuta, Cristina González, Prince Gupta, Jiabo Hu, Yifei Huang, Yiming Huang, Weslie Khoo, Anush Kumar, Robert Kuo, Sach Lakhavani, Miao Liu, Romy Mi Luo, Zhengyi Luo, Brighid Meredith, Austin Miller, Oluwatumininu Oguntola, Xiaqing Pan, Penny Peng, Shraman Pramanick, Merey Ramazanova, Fiona Ryan, W. Shan, Kiran Somasundaram, Chenan Song, Audrey Southerland, Masatoshi Tateno, Huiyu Wang, Yuchen Wang, Takuma Yagi, Mingfei Yan, Xitong Yang, Zecheng Yu, Shengxin Zha, Chen Zhao, Ziwei Zhao, Zhifan Zhu, Jeff Zhuo, Pablo Arbeláez, Gedas Bertasius, David J. Crandall, Dima Damen, Jakob Julian Engel, Giovanni Maria Farinella, Antonino Furnari, Bernard Ghanem, Judy Hoffman, C. V. Jawahar, Richard A. Newcombe, Hyun Soo Park, James M. Rehg, Yoichi Sato, Manolis Savva, Jianbo Shi, Mike Zheng Shou, and Michael Wray. Ego-exo4d: Understanding skilled human activity from first- and third-person perspectives. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19383–19400, 2023. URL https://api.semanticscholar.org/CorpusID:265506384.

Hang Hua, Yunlong Tang, Ziyun Zeng, Liangliang Cao, Zhengyuan Yang, Hangfeng He, Chenliang Xu, and Jiebo Luo. Mmcomposition: Revisiting the compositionality of pre-trained vision-language models. *ArXiv*, abs/2410.09733, 2024. URL https://api.semanticscholar.org/CorpusID:273346558.

Irene Huang, Wei Lin, Muhammad Jehanzeb Mirza, Jacob Hansen, Sivan Doveh, Victor Ion Butoi, Roei Herzig, Assaf Arbelle, Hilde Kuhene, Trevor Darrel, Chuang Gan, Aude Oliva, Rogério Feris, and Leonid Karlinsky. Conme: Rethinking evaluation of compositional reasoning for modern vlms. *ArXiv*, abs/2406.08164, 2024a. URL https://api.semanticscholar.org/CorpusID:270391286.

Yifei Huang, Guo Chen, Jilan Xu, Mingfang Zhang, Lijin Yang, Baoqi Pei, Hongjie Zhang, Lu Dong, Yali Wang, Limin Wang, and Yu Qiao. Egoexolearn: A dataset for bridging asynchronous ego- and exo-centric view of procedural activities in real world. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22072–22086, 2024b. URL https://api.semanticscholar.org/CorpusID:268681223.

OpenAI Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mkadry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alexander Kirillov, Alex Nichol, Alex Paino, Alex Renzin, Alexandre Passos, Alexander Kirillov, Alexi Christakis, Alexis Conneau, Ali Kamali, Allan Jabri, Allison Moyer, Allison Tam, Amadou Crookes, Amin Tootoochian, Amin Tootoonchian, Ananya Kumar, Andrea Vallone, Andrej Karpathy, Andrew Braunstein, Andrew Cann, Andrew Codispoti, Andrew Galu, Andrew Kondrich, Andrew Tulloch, An drey Mishchenko, Angela Baek, Angela Jiang, An toine Pelisse, Antonia Woodford, Anuj Gosalia, Arka Dhar, Ashley Pantuliano, Avi Nayak, Avital Oliver, Barret Zoph, B. Ghorbani, Ben Leimberger, Ben Rossen, Benjamin Sokolowsky, Ben Wang, Benjamin Zweig, Beth Hoover, Blake Samic, Bob McGrew, Bobby Spero, Bogo Giertler, Bowen Cheng, Brad Lightcap, Brandon Walkin, Brendan Quinn, Brian Guarraci, Brian Hsu, Bright Kellogg, Brydon Eastman, Camillo Lugaresi, Carroll L. Wainwright, Cary Bassin, Cary Hudson, Casey Chu, Chad Nelson, Chak Li, Chan Jun Shern, Channing Conger, Charlotte Barette, Chelsea Voss, Chen Ding, Cheng Lu, Chong Zhang, Chris Beaumont, Chris Hallacy, Chris Koch, Christian Gibson, Christina Kim, Christine Choi, Christine McLeavey, Chris Hesse, Claudia Fischer, Clemens Winter, Coley Czarnecki, Colin Jarvis, Colin Wei, Constantin Koumouzelis, Dane Sherburn, Daniel Kappler, Daniel Levin, Daniel Levy, David Carr, David Farhi, David Mély, David Robinson, David Sasaki, Denny Jin, Dev Valladares, Dimitris Tsipras, Doug Li, Phong Duc Nguyen, Duncan Findlay, Edede Oiwoh, Edmund Wong, Ehsan Asdar, Elizabeth Proehl, Elizabeth Yang, Eric Antonow, Eric Kramer, Eric Peterson, Eric Sigler, Eric Wallace, Eugene Brevdo, Evan Mays, Farzad Khorasani, Felipe Petroski Such, Filippo Raso, Francis Zhang, Fred von Lohmann, Freddie Sulit, Gabriel Goh, Gene Oden, Geoff Salmon, Giulio Starace, Greg Brockman, Hadi Salman, Hai-Biao Bao, Haitang Hu, Hannah Wong, Haoyu Wang, Heather Schmidt, Heather Whitney, Hee woo Jun, Hendrik Kirchner, Henrique Pondé de Oliveira Pinto, Hongyu Ren, Huiwen Chang, Hyung Won Chung, Ian Kivlichan, Ian O'Connell, Ian Osband, Ian Silber, Ian Sohl, Ibrahim Okuyucu, Ikai Lan, Ilya Kostrikov, Ilya Sutskever, Ingmar Kanitscheider, Ishaan Gulrajani, Jacob Coxon, Jacob Menick, Jakub W. Pachocki, James Aung, James Betker, James Crooks, James Lennon, Jamie Ryan Kiros, Jan Leike, Jane Park, Jason Kwon, Jason Phang, Jason Teplitz, Jason Wei, Jason Wolfe, Jay Chen, Jeff Harris, Jenia Varavva, Jessica Gan Lee, Jessica Shieh, Ji Lin, Jiahui Yu, Jiayi Weng, Jie Tang, Jieqi Yu, Joanne Jang, Joaquin Quiñonero Candela, Joe Beutler, Joe Landers, Joel Parish, Johannes Heidecke, John Schulman, Jonathan Lachman, Jonathan McKay, Jonathan Uesato, Jonathan Ward, Jong Wook Kim, Joost Huizinga, Jordan Sitkin, Jos Kraaijeveld, Joshua Gross, Josh Kaplan, Josh Snyder, Joshua Achiam, Joy Jiao, Joyce Lee, Juntang Zhuang, Justyn Harriman, Kai Fricke, Kai Hayashi, Karan Singhal, Katy Shi, Kavin Karthik, Kayla Wood, Kendra Rimbach, Kenny Hsu, Kenny Nguyen, Keren Gu-Lemberg, Kevin Button, Kevin Liu, Kiel Howe, Krithika Muthukumar, Kyle Luther, Lama Ahmad, Larry Kai, Lauren Itow, Lauren Workman, Leher Pathak, Leo Chen, Li Jing, Lia Guy, Liam Fedus, Liang Zhou, Lien Mamitsuka, Lilian Weng, Lindsay McCallum, Lindsey Held, Ouyang Long, Louis Feuvrier, Lu Zhang, Lukasz Kondraciuk, Lukasz Kaiser, Luke Hewitt, Luke Metz, Lyric Doshi, Mada Aflak, Maddie Simens, Made laine Boyd, Madeleine Thompson, Marat Dukhan, Mark Chen, Mark Gray, Mark Hudnall, Marvin Zhang, Marwan Aljubeh, Ma teusz Litwin, Matthew Zeng, Max Johnson, Maya Shetty, Mayank Gupta, Meghan Shah, Mehmet Ali Yatbaz, Mengxue Yang, Mengchao Zhong, Mia Glaese, Mianna Chen, Michael Janner, Michael

Lampe, Michael Petrov, Michael Wu, Michele Wang, Michelle Fradin, Michelle Pokrass, Miguel Castro, Miguel Castro, Mikhail Pavlov, Miles Brundage, Miles Wang, Mina Khan, Mira Murati, Mo Bavarian, Molly Lin, Murat Yesildal, Nacho Soto, Natalia Gimelshein, Na talie Cone, Natalie Staudacher, Natalie Summers, Natan LaFontaine, Neil Chowdhury, Nick Ryder, Nick Stathas, Nick Turley, Nikolas A. Tezak, Niko Felix, Nithanth Kudige, Nitish Shirish Keskar, Noah Deutsch, Noel Bundick, Nora Puckett, Ofir Nachum, Ola Okelola, Oleg Boiko, Oleg Murk, Oliver Jaffe, Olivia Watkins, Olivier Godement, Owen Campbell-Moore, Patrick Chao, Paul McMillan, Pavel Belov, Peng Su, Peter Bak, Peter Bakkum, Peter Deng, Peter Dolan, Peter Hoeschele, Peter Welinder, Phil Tillet, Philip Pronin, Phil Tillet, Prafulla Dhariwal, Qim ing Yuan, Rachel Dias, Rachel Lim, Rahul Arora, Rajan Troll, Randall Lin, Raphael Gontijo Lopes, Raul Puri, Reah Miyara, Reimar H. Leike, Renaud Gaubert, Reza Zamani, Ricky Wang, Rob Donnelly, Rob Honsby, Rocky Smith, Rohan Sahai, Rohit Ramchandani, Romain Huet, Rory Carmichael, Rowan Zellers, Roy Chen, Ruby Chen, Ruslan Ramilevich Nigmatullin, Ryan Cheu, Saachi Jain, Sam Altman, Sam Schoenholz, Sam Toizer, Samuel Miserendino, Sandhini Agarwal, Sara Culver, Scott Ethersmith, Scott Gray, Sean Grove, Sean Metzger, Shamez Hermani, Shantanu Jain, Shengjia Zhao, Sherwin Wu, Shino Jomoto, Shirong Wu, Shuaiqi Xia, Sonia Phene, Spencer Papay, Srinivas Narayanan, Steve Coffey, Steve Lee, Stewart Hall, Suchir Balaji, Tal Broda, Tal Stramer, Tao Xu, Tarun Gogineni, Taya Christianson, Ted Sanders, Tejal Patwardhan, Thomas Cunninghman, Thomas Degry, Thomas Dimson, Thomas Raoux, Thomas Shadwell, Tianhao Zheng, Todd Underwood, Todor Markov, Toki Sherbakov, Tom Rubin, Tom Stasi, Tomer Kaftan, Tristan Heywood, Troy Peterson, Tyce Walters, Tyna Eloundou, Valerie Qi, Veit Moeller, Vinnie Monaco, Vishal Kuo, Vlad Fomenko, Wayne Chang, Weiyi Zheng, Wenda Zhou, Wesam Manassra, Will Sheu, Wojciech Zaremba, Yash Patil, Yilei Qian, Yongjik Kim, Youlong Cheng, Yu Zhang, Yuchen He, Yuchen Zhang, Yujia Jin, Yunxing Dai, and Yury Malkov. Gpt-4o system card. *ArXiv*, abs/2410.21276, 2024. URL https://api.semanticscholar.org/CorpusID:273662196.

Mohamed Fazli Mohamed Imam, Chenyang Lyu, and Alham Fikri Aji. Can multimodal llms do visual temporal understanding and reasoning? the answer is no! *ArXiv*, abs/2501.10674, 2025. URL https://api.semanticscholar.org/CorpusID:275758457.

Gemma Team Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ram'e, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean-Bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gael Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Róbert Istvan Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesht Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa Saade, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, Andr'as Gyorgy, André Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia Paterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Boxi Wu, Bobak Shahriari, Bryce Petrini, Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, Cj Carey, Cormac Brick, Daniel Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Paparas, Divyashree Shivakumar Sreepathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huizenga, Eugene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron, Hadi Hashemi, Hanna Klimczak-Pluci'nska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hussein Hazimeh, Ian Ballantyne, Idan Szpektor, Ivan Nardini, Jean Pouget-Abadie, Jetha Chan, Joe Stanton, J. Michael Wieting, Jonathan Lai, Jordi Orbay, Joe Fernandez, Joshua Newlan, Junsong Ji, Jyotinder Singh, Kat Black, Kathy Yu, Kevin Hui, Kiran Vodrahalli, Klaus Greff, Linhai Qiu, Marcella Valentine, Marina Coelho, Marvin Ritter, Matt Hoffman, Matthew Watson, Mayank Chaturvedi, Michael Moynihan, Min Ma, Nabila Babar, Natasha Noy, Nathan Byrd, Nick Roy, Nikola Momchev, Nilay Chauhan, Oskar Bunyan, Pankil Botarda, Paul Caron, Paul Kishan Rubenstein, Phil Culliton, Philipp Schmid, Pier Giuseppe Sessa, Pingmei Xu, Piotr Stańczyk, Pouya Dehghani Tafti, Rakesh Shivanna, Renjie Wu, Renke Pan, Reza Ardeshir Rokni, Rob Willoughby, Rohith Vallu, Ryan Mullins, Sammy Jerome, Sara Smoot, Sertan Girgin, Shariq Iqbal, Shashir Reddy, Shruti Sheth, Siim Põder, Sijal Bhatnagar, Sindhu Raghuram Panyam, Sivan Eiger, Susan Zhang, Tianqi Liu, Trevor Yacovone, Tyler Liechty, Uday Kalra, Utku Evci, Vedant Misra, Vincent Roseberry, Vladimir Feinberg, Vlad Kolesnikov, Woohyun Han, Woosuk Kwon, Xi Chen, Yinlam Chow, Yuvein Zhu, Zichuan Wei, Zoltan Egyed, Victor Cotruta, Minh Giang, Phoebe Kirk, Anand Rao, Jessica Lo, Erica Moreira, Luiz Gustavo

Martins, Omar Sanseviero, Lucas Gonzalez, Zach Gleicher, Tris Warkentin, Vahab S. Mirrokni, Evan Senter, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, Yossi Matias, D. Sculley, Slav Petrov, Noah Fiedel, Noam Shazeer, Oriol Vinyals, Jeffrey Dean, Demis Hassabis, Koray Kavukcuoglu, Clément Farabet, Elena Buchatskaya, Jean-Baptiste Alayrac, Rohan Anil, Dmitry Lepikhin, Sebastian Borgeaud, Olivier Bachem, Armand Joulin, Alek Andreev, Cassidy Hardin, Robert Dadashi, and L'eonard Hussenot. Gemma 3 technical report. *ArXiv*, abs/2503.19786, 2025. URL https://api.semanticscholar.org/CorpusID:277313563.

Ouguzhan Fatih Kar, Alessio Tonioni, Petra Poklukar, Achin Kulshrestha, Amir Zamir, and Federico Tombari. Brave: Broadening the visual encoding of vision-language models. In *European Conference on Computer Vision*, 2024. URL https://api.semanticscholar.org/CorpusID:269033274.

Fanheng Kong, Jingyuan Zhang, Hongzhi Zhang, Shi Feng, Daling Wang, Linhao Yu, Xingguang Ji, Yu Tian, Qi Wang, Fuzheng Zhang, Elmira Amirloo, Jean-Philippe Fauconnier, Christoph Roesmann, Christian Kerl, Rinu Boney, Yu-Mei Qian, Zirui Wang, Afshin Dehghan, Zhe Yinfei Yang, Hidehisa Arai, Keita Miwa, Kento Sasaki, Yu Ya-maguchi, Kohei Watanabe, Shunsuke Aoki, Issei Yamamoto, Covla, Davide Caffagni, Federico Cocchi, Luca Barsellotti, Nicholas Moratelli, Sara Sarto, Lorenzo Baraldi, Marcella Cornia, Rita Cuc-chiara, Zhe Chen, Weiyun Wang, Haowen Tian, Sheng-Tao Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, Bo Li, Yuan hang Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, Wei Li, Zejun Ma, Chunyuan Li, Junnan Li, Dongxu Li, Silvio Savarese, Kunchang Li, Yali Wang, Yi-Wei He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Ye Liu, Zongyang Ma, Zhongang Qi, Yang Wu, and Changzheng Chen. Tuna: Comprehensive fine-grained temporal understanding evaluation on dense dynamic videos. In *Annual Meeting of the Association for Computational Linguistics*, 2025. URL https://api.semanticscholar.org/CorpusID:278911419.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, Limin Wang, and Yu Qiao. Mvbench: A comprehensive multi-modal video understanding benchmark. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22195–22206, 2023. URL https://api.semanticscholar.org/CorpusID:265466214.

Yijun Liang, Ming Li, Chenrui Fan, Ziyue Li, Dang Nguyen, K.A. Cobbina, Shweta Bhardwaj, Jiuhai Chen, Fuxiao Liu, and Tianyi Zhou. Colorbench: Can vlms see and understand the colorful world? a comprehensive benchmark for color perception, reasoning, and robustness. *ArXiv*, abs/2504.10514, 2025. URL https://api.semanticscholar.org/CorpusID:277786803.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *ArXiv*, abs/2304.08485, 2023. URL https://api.semanticscholar.org/CorpusID:258179774.

Ye Liu, Zongyang Ma, Zhongang Qi, Yang Wu, Ying Shan, and Chang Wen Chen. E.t. bench: Towards open-ended event-level video-language understanding. *ArXiv*, abs/2409.18111, 2024a. URL https://api.semanticscholar.org/CorpusID:272911111.

Yuanxin Liu, Shicheng Li, Yi Liu, Yuxiang Wang, Shuhuai Ren, Lei Li, Sishuo Chen, Xu Sun, and Lu Hou. Tempcompass: Do video llms really understand videos? In *Annual Meeting of the Association for Computational Linguistics*, 2024b. URL https://api.semanticscholar.org/CorpusID:268201547.

Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. Egoschema: A diagnostic benchmark for very long-form video language understanding. *ArXiv*, abs/2308.09126, 2023. URL https://api.semanticscholar.org/CorpusID:261031047.

Jong Sung Park, Kanchana Ranasinghe, Kumara Kahatapitiya, Wonjeong Ryoo, Donghyun Kim, and Michael S. Ryoo. Too many frames, not all useful: Efficient strategies for long-form video qa. *ArXiv*, abs/2406.09396, 2024. URL https://api.semanticscholar.org/CorpusID:270440923.

Chiara Plizzari, Alessio Tonioni, Yongqin Xian, Achin Kulshrestha, and Federico Tombari. Omnia de egotempo: Benchmarking temporal understanding of multi-modal llms in egocentric videos. *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24129–24138, 2025. URL https://api.semanticscholar.org/CorpusID:277104551.

Ruchit Rawal, Khalid Saifullah, Ronen Basri, David Jacobs, Gowthami Somepalli, and Tom Goldstein. Cinepile: A long video question answering dataset and benchmark. *ArXiv*, abs/2405.08813, 2024. URL https://api.semanticscholar.org/CorpusID:269761335.

Ziyao Shangguan, Chuhan Li, Yuxuan Ding, Yanan Zheng, Yilun Zhao, Tesca Fitzgerald, and Arman Cohan. Tomato: Assessing visual temporal reasoning capabilities in multimodal foundation models. *ArXiv*, abs/2410.23266, 2024. URL https://api.semanticscholar.org/CorpusID:273695606.

Yingjin Song, Yupei Du, Denis Paperno, and Albert Gatt. Burn after reading: Do multimodal large language models truly capture order of events in image sequences? In *Annual Meeting of the Association for Computational Linguistics*, 2025. URL https://api.semanticscholar.org/CorpusID:279319023.

Yanpeng Sun, Huaxin Zhang, Qiang Chen, Xinyu Zhang, Nong Sang, Gang Zhang, Jingdong Wang, and Zechao Li. Improving multi-modal large language model through boosting vision capabilities. *ArXiv*, abs/2410.13733, 2024. URL https://api.semanticscholar.org/CorpusID:273404344.

Sirnam Swetha, Hildegard Kuehne, and Mubarak Shah. Timelogic: A temporal logic benchmark for video qa. *ArXiv*, abs/2501.07214, 2025. URL https://api.semanticscholar.org/CorpusID:275470839.

Jiayu Wang, Yifei Ming, Zhenmei Shi, Vibhav Vineet, Xin Wang, and Neel Joshi. Is a picture worth a thousand words? delving into spatial reasoning for vision language models. *ArXiv*, abs/2406.14852, 2024. URL https://api.semanticscholar.org/CorpusID:270688598.

Xin Wang, Taein Kwon, Mahdi Rad, Bowen Pan, Ishani Chakraborty, Sean Andrist, Dan Bohus, Ashley Feniello, Bugra Tekin, Felipe Vieira Frujeri, Neel Joshi, and Marc Pollefeys. Holoassist: an egocentric human interaction dataset for interactive ai assistants in the real world. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 20213–20224, 2023. URL https://api.semanticscholar.org/CorpusID:263310718.

Zeqing Wang, Shiyuan Zhang, Chengpei Tang, and Keze Wang. Timecausality: Evaluating the causal ability in time dimension for vision language models. *ArXiv*, abs/2505.15435, 2025. URL https://api.semanticscholar.org/CorpusID:278782911.

Junxiao Xue, Quan Deng, Fei Yu, Yanhao Wang, Jun Wang, and Yuehua Li. Enhanced multimodal rag-llm for accurate visual question answering. *ArXiv*, abs/2412.20927, 2024. URL https://api.semanticscholar.org/CorpusID:275133918.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxin Yang, Jingren Zhou, Jingren Zhou, Junyan Lin, Kai Dang, Keqin Bao, Ke-Pei Yang, Le Yu, Li-Chun Deng, Mei Li, Min Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shi-Qiang Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yi-Chao Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report. *ArXiv*, abs/2505.09388, 2025a. URL https://api.semanticscholar.org/CorpusID:278602855.

Qi Yang, Bolin Ni, Shiming Xiang, Han Hu, Houwen Peng, and Jie Jiang. R-4b: Incentivizing general-purpose auto-thinking capability in mllms via bi-mode annealing and reinforce learning. *ArXiv*, abs/2508.21113, 2025b. URL https://api.semanticscholar.org/CorpusID:280985221.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. *ArXiv*, abs/2308.02490, 2023. URL https://api.semanticscholar.org/CorpusID:260611572.

Yuqian Yuan, Ronghao Dang, Long Li, Wentong Li, Dian Jiao, Xin Li, Deli Zhao, Fan Wang, Wenqiao Zhang, Jun Xiao, and Yueting Zhuang. Eoc-bench: Can mllms identify, recall, and forecast objects in an egocentric world? *ArXiv*, abs/2506.05287, 2025. URL https://api.semanticscholar.org/CorpusID:279244628.

Maxime Zanella and Ismail Ben Ayed. Low-rank few-shot adaptation of vision-language models. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1593–1603, 2024. URL https://api.semanticscholar.org/CorpusID:270094592.

Yunan Zeng, Yan Huang, Jinjin Zhang, Zequn Jie, Zhenhua Chai, and Liang Wang. Investigating compositional challenges in vision-language models for visual grounding. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14141–14151, 2024. URL https://api.semanticscholar.org/CorpusID:272722693.

Jianrui Zhang, Mu Cai, and Yong Jae Lee. Vinoground: Scrutinizing lmms over dense temporal reasoning with short videos. *ArXiv*, abs/2410.02763, 2024. URL https://api.semanticscholar.org/CorpusID:273098539.

Yiyang Zhou, Linjie Li, Shi Qiu, Zhengyuan Yang, Yuyang Zhao, Siwei Han, Yangfan He, Kangqi Li, Haonian Ji, Zihao Zhao, Haibo Tong, Lijuan Wang, and Huaxiu Yao. Glimpse: Do large vision-language models truly think with videos or just glimpse at them? *ArXiv*, abs/2507.09491, 2025. URL https://api.semanticscholar.org/CorpusID:280295497.

A Ablation on Initial Shuffle Order

A.1 Methodology

The goal of this ablation experiment is to assess how sensitive the model's performance is to the initial ordering of input frames in the frame-ordering task. To do so, we constructed a small dataset of 10 medium-length clips, each consisting of 4 frames, and ran Gemini 3.0 Flash on all 24 possible permutations of the frames for each clip using the baseline configuration. For each permutation, we recorded the model's predicted order and computed Kendall's Tau metric relative to the ground truth. By grouping results according to permutation, this analysis quantifies the extent to which current vision-language models are affected by arbitrary input shuffles and highlights the effect the initial shuffle has on a model's benchmark performance.

A.2 Results and Analysis

As shown in Figure 3, results of our ablation study shows that performance varies substantially depending on the input permutation order, with Kendall's τ ranging from -0.233 to 0.933 and an overall mean of 0.378 ($\sigma = 0.274$). The distribution was right-skewed, with a median of 0.400 and an interquartile range of 0.417. Most permutations yielded positive correlations, but two permutations (3,1,4,2 and 4,2,1,3) produced negative correlations, indicating that particular input orderings can lead to performance worse than random chance.

A pronounced first-position anchoring effect was observed. When the correct first frame appeared in the first input position, the model achieved a mean τ of 0.667. Performance declined sharply when Frame 1 was presented in later positions: position 2 (0.283), position 3 (0.267), and position 4 (0.294). This indicates that the model heavily anchors its temporal reasoning on the first frame in the input sequence, often incorrectly assuming it represents the chronological beginning, even when explicitly prompted that frames are out of order. A weaker but similar "last-position anchoring" effect was also evident: performance improved when the correct last frame appeared later in the input sequence, with τ increasing from 0.256 at position 1 to 0.528 at position 4. These effects demonstrate that the model relies on positional cues rather than purely on visual-temporal information.

Performance exhibited a clear negative correlation with input disorder, measured by the number of pairwise inversions in the input sequence. With 0 inversions (correct order 1,2,3,4), the model achieved near-perfect performance with $\tau = 0.933$. Performance declined progressively with more scrambled inputs: 1 inversion (0.600), 2 inversions (0.453), 3 inversions (0.339), and 4 inversions (0.187). Although the relationship was not strictly monotonic—permutations with 5 inversions occasionally outperformed those with 4—the overall trend showed that more scrambled inputs led to worse temporal reconstruction. The fully reversed permutation (4,3,2,1) achieved only $\tau = 0.200$, highlighting the model's reliance on input order as a heuristic rather than genuine visual reasoning.

The top-performing permutations were those that preserved the correct first frame or maintained frames close to their true temporal order, including 1,2,3,4 ($\tau = 0.933$), 1,3,4,2 ($\tau = 0.667$), and 1,4,3,2 ($\tau = 0.667$). Conversely, the worst-performing permutations typically placed later frames first, such as 3,1,4,2 ($\tau = -0.233$) and 4,2,1,3 ($\tau = -0.133$). These results reveal a fundamental limitation of current vision-language models: strong sensitivity to frame position, which can confound evaluations of temporal reasoning.

These findings have important implications for benchmark design and model evaluation. Evaluation protocols must randomize input order, as testing only on correctly-ordered or consistently-shuffled sequences produces misleading performance estimates. Single-permutation evaluations are unreliable; for instance, the model appears near-perfect on the forward permutation but can perform worse than random on other orderings. The first-position anchoring effect further suggests that models may learn spurious correlations between input position and temporal order rather than true temporal reasoning.

To mitigate position bias in our main evaluation, we adopted a balanced sampling approach in which each of the 156 video clips was evaluated across uniformly random permutations, ensuring that every frame had an equal probability of occupying each input position. To maintain consistency across models, we recorded the specific shuffled orders used for each question when evaluating the first model and reused these exact orders for all subsequent models. This procedure ensures that differences in performance reflect genuine model variation rather than inconsistencies in input ordering. Evaluating

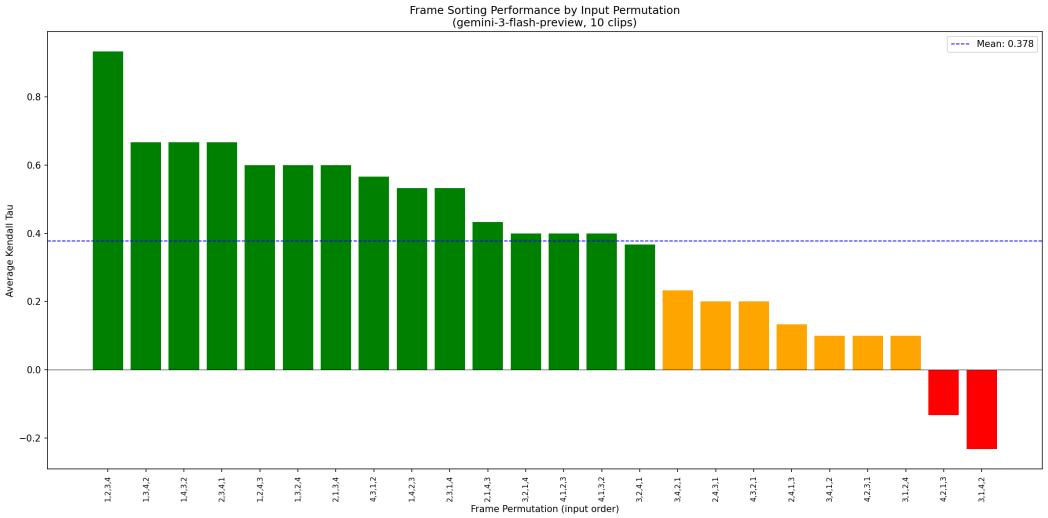

Figure 3: Frame-sorting performance of Gemini 3 Flash across all 24 possible input permutations of 4 frames, evaluated on 10 video clips. Kendall's Tau ranges from -0.23 to 0.93. Bars are colored green/orange for positive correlations and red for negative correlations. The dashed blue line at $y = 0.378$ represents the mean Kendall's Tau across all permutations.

a large number of clips with uniformly random permutations averages out positional biases and provides a more accurate estimate of each model's true visual-temporal reasoning capability. This analysis thus serves as a diagnostic tool for quantifying position bias, while our main evaluation methodology explicitly attempts to account for it through balanced sampling and consistent shuffling across models.

B LIMITATIONS

Our study has several notable limitations that suggest directions for future work. First, the fine-tuning experiments were conducted on a relatively small dataset of pairwise frame comparisons, which limited the generalization capacity of the jointly tuned LM + Vision Encoder model and led to overfitting. Expanding the training set with additional egocentric clips or synthetic temporal augmentations could allow more stable joint adaptation and reveal the true extent of how large-scale multimodal tuning affects temporal reasoning. Based on prior studies, we expect that with a sufficiently large dataset, this LM + Vision Encoder fine-tuning strategy would further improve temporal performance beyond what LM-only fine-tuning achieved (Zanella & Ayed, 2024; Sun et al., 2024).

Second, while our random permutation strategy helps mitigate the positional bias revealed in our permutation robustness analysis, it does not fully eliminate it. The pronounced order sensitivity observed in our experiments suggests that more principled bias correction such as adversarial permutation testing, order-invariant representations, or explicit positional debiasing during training remains an open challenge. Prompt design also had a strong influence on model performance, with open-source models frequently struggling under complex instructions, particularly in hint-based experiments. Although our study provides initial insights, we acknowledge that the statistical rigor is limited; future work could expand on this by examining variance across multiple random seeds or analyzing confidence intervals to better quantify model reliability. In addition, future studies could also more carefully investigate the effects of prompts, exploring why models are so sensitive to prompt design and how different formulations impact performance.

Finally, computational and financial constraints prevented us from evaluating the most recent or largest proprietary and open-source models, limiting the completeness of our cross-model comparison. Extending this benchmark to state-of-the-art frontier models would provide a clearer understanding of scaling effects and architectural trends in temporal reasoning. Additionally, performing fine-tuning

experiments on a broader set of models could further validate the effectiveness of the two LoRA tuning strategies and provide stronger evidence of its benefits.

C FINE-TUNE DIAGRAM

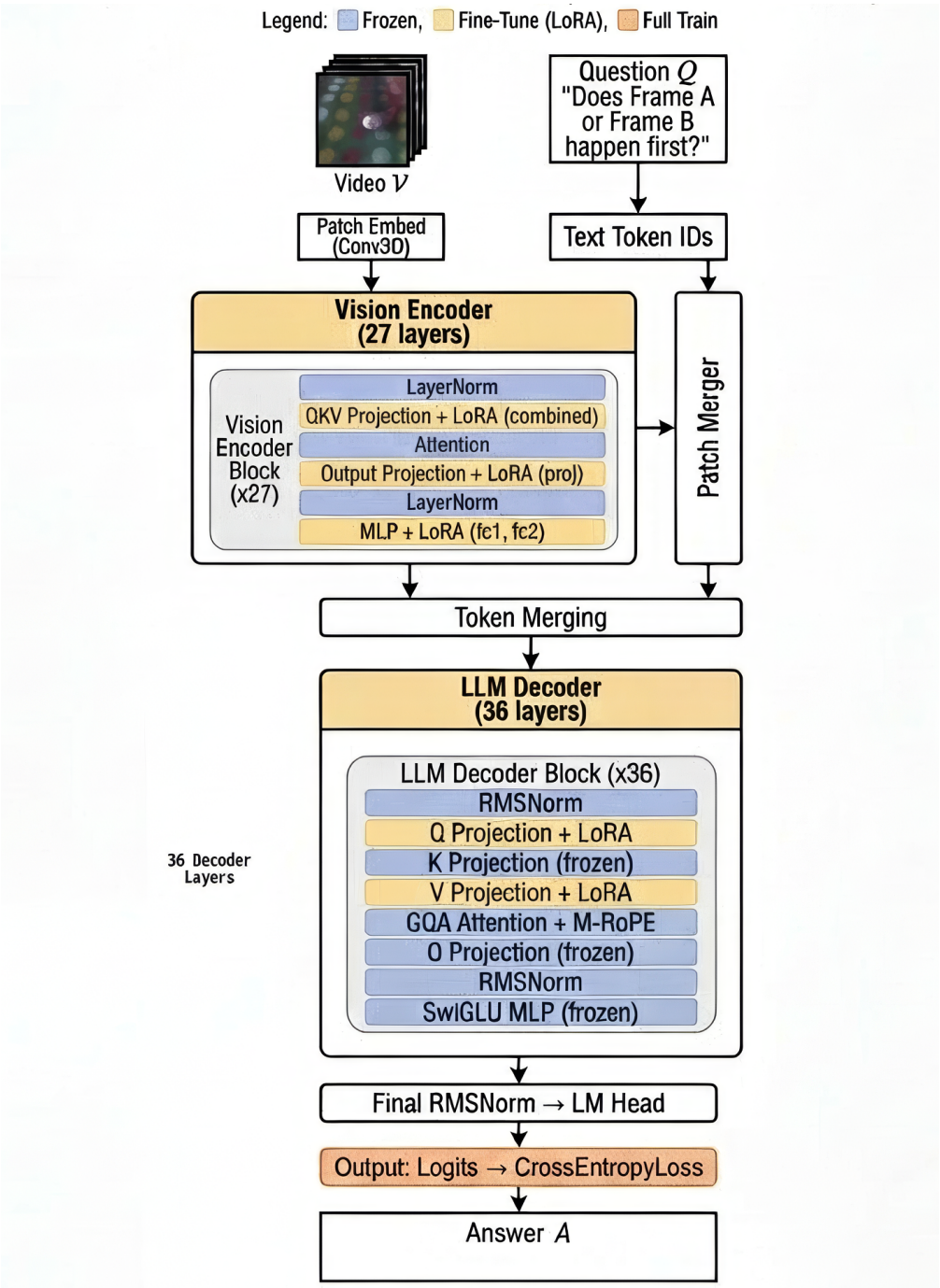

Figure 4: Architecture of the joint language–vision fine-tuning approach. LoRA adapters are applied to both the LLM (q and v projection layers) and the vision encoder (qkv, proj, fc1, fc2 layers), allowing the vision encoder to learn feature representations optimized for temporal ordering while the LLM simultaneously learns to interpret these evolving features.

D RESULTS FOR OTHER METRICS

Table 3: Performance of all evaluated models across nine frame-sorting experiments, reported using the Pairwise Accuracy metric.

Experiment	GPT-4o	Claude-Sonnet-4.5	Gemini-3.0-Flash	Gemma 4B	R-4B	Phi 3.5 Vision	Llava-13B	Qwen3-VL-8B	Qwen3-VL-32B	Human Baseline
Original	0.61	0.54	**0.67**	0.49	0.52	0.49	0.53	0.57	0.56	0.77
Changing Frames										
2 Frames	0.65	**0.67**	**0.67**	0.63	0.52	0.46	0.51	0.54	0.59	0.87
8 Frames	0.56	0.54	**0.65**	0.50	0.50	0.50	0.52	0.53	0.59	0.68
Changing Video Length										
Short	0.58	0.55	**0.63**	0.48	0.53	0.48	0.52	0.55	0.55	0.70
Long	0.53	0.56	**0.65**	0.56	0.48	0.50	0.49	0.58	0.59	0.72
Changing Annotation Description										
None	0.60	0.52	**0.65**	0.51	0.50	0.50	0.50	0.53	0.59	0.74
Comprehensive	0.63	0.58	**0.72**	0.50	0.51	0.53	0.54	0.55	0.56	0.78
Changing Hints										
Frame 1 & 3	0.64	0.75	**0.82**	0.56	0.49	0.50	0.46	0.67	0.73	0.89
Frame 2 & 4	0.73	0.65	**0.80**	0.49	0.49	0.55	0.55	0.71	0.74	0.89

Table 4: Performance of all evaluated models across nine frame-sorting experiments, reported using the Mean Absolute Distance (MAD) metric. Lower MAD values indicate better performance.

Experiment	GPT-4o	Claude-Sonnet-4.5	Gemini-3.0-Flash	Gemma 4B	R-4B	Phi 3.5 Vision	Llava-13B	Qwen3-VL-8B	Qwen3-VL-32B	Human Baseline
Original	0.99	1.13	**0.80**	1.25	1.20	1.26	1.19	1.05	1.08	0.56
Changing Frames										
2 Frames	0.35	**0.33**	**0.33**	0.41	0.48	0.54	0.49	0.46	0.41	0.13
8 Frames	2.34	2.35	**1.90**	2.69	2.61	2.68	2.50	2.42	2.14	1.70
Changing Video Length										
Short	1.06	1.10	**0.88**	1.34	1.20	1.31	1.23	1.14	1.08	0.75
Long	1.51	1.07	**0.87**	1.08	1.29	1.27	1.29	1.06	1.01	0.97
Changing Annotation Description										
None	1.00	1.20	**0.91**	1.25	1.22	1.21	1.26	1.13	1.05	0.63
Comprehensive	0.93	1.04	**0.71**	1.32	1.20	1.19	1.21	1.07	1.06	0.59
Changing Hints										
Frame 1 & 3	0.36	0.25	**0.18**	0.44	0.51	0.50	0.54	0.33	0.27	0.23
Frame 2 & 4	0.27	0.35	**0.20**	0.51	0.51	0.61	0.45	0.29	0.26	0.25

Table 5: Performance of all evaluated models across nine frame-sorting experiments, reported using the Longest Common Subsequence Ratio (LCSR) metric.

Experiment	GPT-4o	Claude-Sonnet-4.5	Gemini-3.0-Flash	Gemma 4B	R-4B	Phi 3.5 Vision	Llava-13B	Qwen3-VL-8B	Qwen3-VL-32B	Human Baseline
Original	0.68	0.62	**0.72**	0.60	0.61	0.60	0.63	0.66	0.66	0.79
Changing Frames										
2 Frames	0.82	0.83	**0.84**	0.80	0.76	0.73	0.75	0.77	0.79	0.94
8 Frames	0.54	0.51	**0.60**	0.46	0.46	0.47	0.49	0.51	0.51	0.61
Changing Video Length										
Short	0.66	0.65	**0.72**	0.59	0.62	0.59	0.62	0.64	0.67	0.75
Long	0.55	0.64	**0.72**	0.63	0.59	0.61	0.61	0.64	0.67	0.76
Changing Annotation Description										
None	0.68	0.63	**0.72**	0.59	0.63	0.61	0.61	0.64	0.67	0.79
Comprehensive	0.70	0.67	**0.74**	0.59	0.62	0.63	0.63	0.64	0.67	0.73
Changing Hints										
Frame 1 & 3	0.82	0.87	**0.91**	0.78	0.75	0.75	0.73	0.83	0.86	0.90
Frame 2 & 4	0.86	0.83	**0.90**	0.75	0.75	0.70	0.77	0.85	0.87	0.89

Table 6: Performance of all evaluated models across nine frame-sorting experiments, reported using the Edit Distance metric. Lower values indicate better performance.

Experiment	GPT-4o	Claude-Sonnet-4.5	Gemini-3.0-Flash	Gemma 4B	R-4B	Phi 3.5 Vision	Llava-13B	Qwen3-VL-8B	Qwen3-VL-32B	Human Baseline
Original	2.15	2.47	**1.78**	2.64	2.53	2.68	2.48	2.20	2.27	1.32
Changing Frames										
2 Frames	0.71	0.67	**0.65**	0.75	0.96	1.09	0.99	0.92	0.82	0.25
8 Frames	5.44	5.74	**4.90**	6.32	6.57	6.54	6.15	5.87	5.70	5.18
Changing Video Length										
Short	2.15	2.25	**1.78**	2.85	2.57	2.74	2.64	2.36	2.08	1.65
Long	2.46	2.28	**1.95**	2.45	2.76	2.61	2.68	2.30	2.10	1.66
Changing Annotation Description										
None	2.15	2.49	**1.94**	2.69	2.47	2.55	2.66	2.32	2.24	1.35
Comprehensive	2.00	2.22	**1.76**	2.73	2.47	2.46	2.68	2.23	2.14	1.36
Changing Hints										
Frame 1 & 3	0.73	0.50	**0.35**	0.88	1.01	1.00	1.08	0.67	0.54	0.53
Frame 2 & 4	0.56	0.69	**0.39**	1.02	1.02	0.92	0.90	0.59	0.52	0.67

Table 7: Performance of all evaluated models across nine frame-sorting experiments, reported using the Exact Match Rate metric.

Experiment	GPT-4o	Claude-Sonnet-4.5	Gemini-3.0-Flash	Gemma R-4B 4B		Phi 3.5 Vision	Llava-13B	Qwen3-VL-8B	Qwen3-VL-32B	Human Baseline
Original	0.14	0.06	**0.28**	0.05	0.02	0.03	0.06	0.13	0.12	0.48
Changing Frames										
2 Frames	0.65	**0.67**	**0.67**	0.63	0.52	0.46	0.51	0.54	0.59	0.87
8 Frames	0.01	0.00	**0.05**	0.00	0.00	0.00	0.00	0.01	0.00	0.03
Changing Video Length										
Short	0.14	0.12	**0.31**	0.02	0.03	0.02	0.04	0.10	0.20	0.35
Long	0.14	0.08	**0.25**	0.10	0.03	0.04	0.00	0.11	0.16	0.31
Changing Annotation Description										
None	0.15	0.08	**0.21**	0.03	0.07	0.04	0.05	0.09	0.12	0.48
Comprehensive	0.21	0.16	**0.24**	0.02	0.03	0.11	0.08	0.14	0.17	0.44
Changing Hints										
Frame 1 & 3	0.63	0.75	**0.82**	0.56	0.49	0.50	0.46	0.67	0.73	0.74
Frame 2 & 4	0.71	0.65	**0.80**	0.49	0.49	0.53	0.54	0.71	0.74	0.67

E EXPERIMENTS RAN

Table 8: Experimental design showing the nine experiments conducted and the values of each experimental variable.

Experiment	Number of Frames	Video Length	Annotation Level	Hint Level
1	2	Medium	Some	No Hints
2	4	Medium	Some	No Hints
3	8	Medium	Some	No Hints
4	4	Short	Some	No Hints
5	4	Long	Some	No Hints
6	4	Medium	None	No Hints
7	4	Medium	Comprehensive	No Hints
8	4	Medium	Some	Frame 1 & 3
9	4	Medium	Some	Frame 2 & 4

F BRIEF AND COMPREHENSIVE ANNOTATION EXAMPLES

F.1 BRIEF ANNOTATIONS

For the top action of Figure 2:

```
"A person uses a spatula to transfer cooked food pieces from a
skillet into a metal tray."
```

For the middle action of Figure 2:

```
"A person is restocking a vending machine with bags of jerky."
```

For the bottom action of Figure 2:

```
"A person is shown folding a black and white striped
long-sleeved shirt on a bed"
```

F.2 COMPREHENSIVE ANNOTATIONS

For the top action of Figure 2:

```
"In this video, a person is seen cooking in a kitchen.  The
individual is actively engaged in transferring cooked food from
a stainless steel skillet to a rectangular metal tray.  The
person uses a metal spatula to scrape and lift the meat pieces
out of the skillet, which is positioned on a modern-looking gas
stove.  The tray, already partially filled with similar food
pieces, sits on a wooden countertop next to the stove.  To the
left of the tray, there's a stainless steel pot containing
some green leafy vegetables, possibly cabbage.  The person
carefully moves each piece, ensuring the skillet is emptied.
The overall scene is captured from a top-down, slightly angled
perspective, highlighting the food preparation process in a
well-lit kitchen."
```

For the middle action of Figure 2:

```
"The video starts with a close-up of a person's hands as
they restock a vending machine.  The person is seen placing
individual snack bags into the metal coils of the machine.  The
first snack being placed is a white and orange bag, likely
containing dried jerky.  They carefully slide each bag into
the slots between the spiral wires.  After filling one slot,
they pick up another identical bag from a pile below and repeat
the process for the next slot in the row.  The person continues
to fill the row with the white and orange bags, one by one,
ensuring each is properly seated in the coil mechanism.  The
lighting inside the machine is bright.  The video captures the
repetitive and precise nature of restocking a vending machine."
```

For the bottom action of Figure 2:

```
"The video captures a first-person perspective of a person
folding a black and white striped long-sleeved shirt on a white
bed.  The process begins with the shirt laid out flat.  The
person first folds the right side of the shirt towards the
middle, smoothing it down.  Next, they repeat the action with
the left side, bringing it towards the center to form a long,
narrow rectangle.  After smoothing the fabric again, they fold
the bottom third of the shirt upwards, followed by another fold
to the top, resulting in a compact, square shape.  Finally, they
pick up the neatly folded shirt and place it to the side next
to other items of clothing, including a white piece and a red
floral patterned garment."
```

G PROMPTS

PROPRIETARY MODEL WITHOUT HINTS

```
You are shown [total frames] frames from a video.  These frames
have been shuffled and are NOT in their original order.  The
labels "Frame 1", "Frame 2", etc.  Refer to the order they
appear in this message, not chronologically.

The video shows:  [scene context]

Your task:  Determine the correct chronological order of these
frames based on the visual content.

    1. Briefly describe what you observe in each frame.
    2. Explain your reasoning for the temporal order based on:
        • Object positions and movements
        • Progress of any actions being performed
        • Any other visual cues that indicate sequence
    3. Finally, provide your answer in this format:

The correct temporal order is:  [comma-separated frame numbers]

For example (with 8 frames):  "The correct temporal order is:
5, 2, 8, 1, 4, 7, 3, 6"
```

PROPRIETARY MODEL WITH HINTS

```
You are shown [total frames] frames from a video.  These frames
have been shuffled and are NOT in their original order.  The
labels "Frame 1", "Frame 2", etc.  refer to the order they
appear in this message, not their chronological order.

HINTS PROVIDED: [hint frames description]

CRITICAL HINT INFORMATION: The frames marked as HINTS above are
shown to you with their CORRECT temporal positions explicitly
stated.  These hint frames MUST remain in their specified
positions in your final answer.  For example, if a hint says
"Frame X should be in temporal position Y", that means in the
final temporal sequence, Frame X MUST be at position Y -- you
cannot move it to a different position.  Use these fixed hint
frames as anchor points to determine where the remaining frames
should go.

The video shows:  [scene context]

Your task:  Determine the correct chronological order of these
frames based on the visual content.

    1. Briefly describe what you observe in each frame.
    2. Explain your reasoning for the temporal order based on:
        • Object positions and movements
        • Progress of any actions being performed
        • Any other visual cues that indicate sequence
    3. Use the fixed hint frames as anchor points to determine
       where the remaining (non-hint) frames should be placed.
    4. Finally, provide your answer in this format:

The correct temporal order is:  [comma-separated frame numbers]

For example (with 8 frames):  "The correct temporal order is:
5, 2, 8, 1, 4, 7, 3, 6"
```

OPEN-SOURCE MODEL WITHOUT HINTS

```
You are shown [total frames] frames from a video.  These frames
have been shuffled and are NOT in their original order.  The
labels "Frame 1", "Frame 2", etc.  refer to the order they
appear in this message, not their chronological order.

The video shows:  [scene context]

Your task:  Determine the correct chronological order of these
frames based on the visual content.

First, briefly describe what you observe in each frame.  Then
explain your reasoning for the temporal order based on:

    • Object positions and movements

    • Progress of any actions being performed

    • Any other visual cues that indicate sequence

Finally, provide your answer in this format:
"The correct temporal order is:  [comma-separated frame
numbers]"

For example (with 8 frames):  "The correct temporal order is:
5, 2, 8, 1, 4, 7, 3, 6"
```

OPEN-SOURCE MODEL PROMPT WITH HINTS

The hint-based experiments involved two possible answer options, as two of the four frame positions were already provided as reference points. For open-source models, we explicitly specified these two valid answer options to focus evaluation on their ability to infer the relative order of the remaining frames, rather than on interpreting longer, multi-step instructions. Proprietary models, by contrast, were able to handle open-ended prompts reliably, so we preserved the original format.

```
You are shown [total_frames] frames from a video.  These frames
have been shuffled and are NOT in their original order.  The
labels "Frame 1", "Frame 2", etc.  refer to the order they
appear in this message, not their chronological order.

The video shows:  [scene context]

GUARANTEED CORRECT HINTS (you MUST use these exactly):
[hint_frames_description]

Your task:
    1. Look at Frame [non_hint a] and Frame [non_hint b] -- these
       are the only frames you need to decide about.
    2. Describe what you see in each of the four frames.
    3. Decide:  which one ([non_hint a] or [non_hint b])
       happened 1st chronologically?  Which one happened 3rd
       chronologically?

CRITICAL: Your final answer MUST have EXACTLY 4 numbers
separated by commas.
The format is:  The correct temporal order is:  X, [hint_pos2],
Y, [hint_pos4]
Where X must be either [non_hint a] or [non_hint b], and Y must
be the other one.

There are ONLY 2 valid answers:
    • The correct temporal order is:  [non_hint a], [hint_pos2],
      [non_hint b], [hint_pos4]
    • The correct temporal order is:  [non_hint b], [hint_pos2],
      [non_hint a], [hint_pos4]

You MUST choose one of these two options.
```

H ADDITIONAL DATASET INFORMATION

Table 9: Summary of our egocentric benchmark dataset. Counts and distributions of key properties are provided along with examples.

Category / Property	**Count / Description**	**Examples**
Total Clips	1,056	Continuous egocentric recordings spanning diverse everyday activities
Action Categories	58	Pouring, folding, stacking, opening containers, cleaning, etc.
Environments	14	Kitchen, office, living room, backyard, warehouse, etc.
Participants	1–2 per clip	All adults; mixed genders
Occlusion Characteristics	Varies	Hand or utensil partially occludes objects; occasional objects out of frame due to camera shifts
Motion Characteristics	Varies	Fast, slow, repetitive, abrupt, or subtle hand/object movements

Table 10: Summary of our egocentric LoRA training dataset. Counts and distributions of key properties are provided along with examples.

Category / Property	**Count / Description**	**Examples**
Total Clips	421	Egocentric recordings collected for LoRA fine-tuning experiments
Action Categories	17	Bathing, tidying, packing, ironing, decluttering, unboxing, etc.
Environments	12	Kitchen, bathroom, library, living room, outdoors, etc.
Participants	1–2 per clip	Mostly adults; mixed genders
Occlusion Characteristics	Varies	Hand or utensil partially occludes objects; occasional objects out of frame
Motion Characteristics	Varies	Fast, slow, repetitive, abrupt, or subtle hand/object movements

I CODE AND DATASET AVAILABILITY

Code Repository: `https://anonymous.4open.science/r/EGO-FLIGHT-37BF/`

Partial Dataset: `https://zenodo.org/records/18407442`

The supplementary dataset includes a subset of the full EGO-FLIGHT dataset consisting of 256 short videos. The full dataset will be publicly released upon acceptance of the paper.