

TEMPURA: Temporal Event Masked Prediction and Understanding for Reasoning in Action

Jen-Hao Cheng¹, Yi-Hao Peng², Huapeng Zhou¹, Vivian Wang¹, Huayu Wang¹,
Hsiang-Wei Huang¹, Wenhao Chai³, Hou-I Liu⁴, Kuang-Ming Chen¹, Cheng-Yen Yang¹,
Yi-Ling Chen⁵, Vibhav Vineet⁵, Qin Cai¹, Jenq-Neng Hwang¹

¹ University of Washington

² Carnegie Mellon University

³ Princeton University

⁴ National Yang Ming Chiao Tung University

⁵ Microsoft

Link: [Project Page](#) | [VER Data](#) | [Code & Model](#)

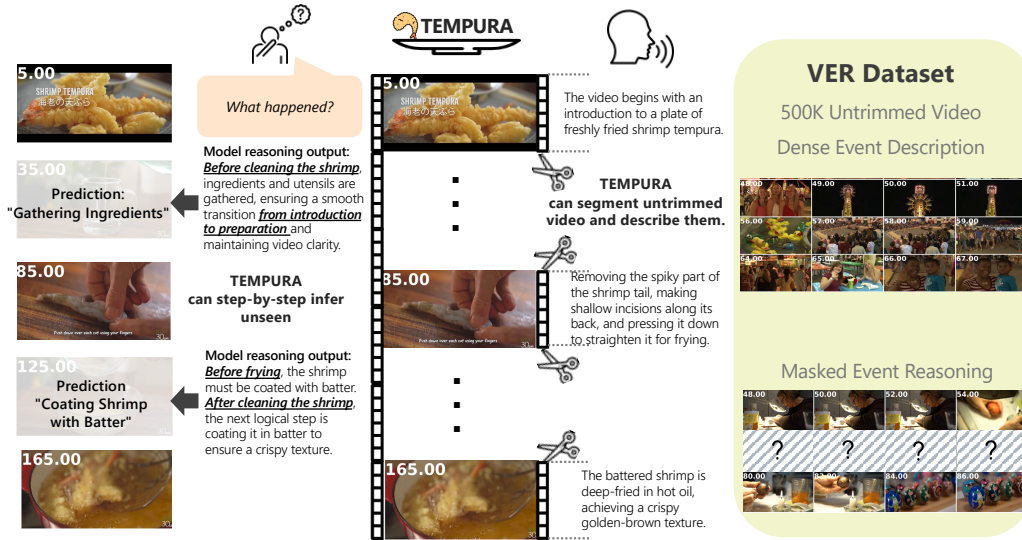


Figure 1: TEMPURA learns video understanding with a two-stage curriculum. It first reconstructs event structure and causal links by filling missing steps in event sequences. It then segments each video into non-overlapping events and writes dense event captions.

Abstract

Understanding causal event relationships and achieving fine-grained temporal grounding in videos remain challenging for vision-language models (VLMs). We propose TEMPURA (Temporal Event Masked Prediction and Understanding for Reasoning in Action), a two-stage training framework that enhances the video temporal understanding of VLMs. Inspired by infilling techniques in language modeling, TEMPURA first performs masked event prediction, learning to reconstruct missing events and generate step-by-step causal explanations from dense event annotations. It then learns video segmentation and dense captioning, decomposing videos into non-overlapping events with detailed, timestamp-aligned descriptions. We train TEMPURA on VER, our large-scale dataset of 500K videos annotated with temporally aligned event descriptions and structured reasoning steps. Experiments on video temporal grounding and highlight detection benchmarks show that TEMPURA substantially improves strong base VLMs across model families and scales, confirming that combining event-level reasoning with fine-grained temporal segmentation is an effective recipe for video temporal understanding.

1 Introduction

Recent video large multi-modal models (LMMs) (Liu et al., 2023; Li et al., 2025a; Maaz et al., 2024; Bai et al., 2025) have extended large language models (LLMs) with video understanding. Yet fine-grained temporal understanding remains difficult. Recent training pipelines improve video instruction following with curated data or preference optimization (Zhang et al., 2025b; Li et al., 2025b), but they still do not explicitly teach models to segment videos into coherent events or infer missing causal steps between events.

We address this gap by introducing **TEMPURA** (*Temporal Event Masked Prediction and Understanding for Reasoning in Action*), a two-stage training pipeline that combines event reasoning with dense temporal grounding. In the first stage, TEMPURA masks an event and trains the model to infer the missing event together with a step-by-step causal explanation. Inspired by Fill-in-the-Middle (FIM) (Bavarian et al., 2022; Shen et al., 2023), we build this supervision by masking dense event captions and asking a strong LLM to generate a plausible pseudo-event and its reasoning trace. In the second stage, the model learns to segment an untrimmed video into non-overlapping events with explicit start and end timestamps and detailed captions. This stage grounds the event structure learned in stage one and teaches the model to cover the full video without specialized temporal heads. Together, the two stages perform event-centric temporal representation learning: the model first captures how events unfold and connect over time, and then grounds this structure with explicit timestamps.

To support this training recipe, we construct **VER**, a dataset built through a multi-step event annotation pipeline (Figure 3). We filter dynamic videos from YT-1B (Zellers et al., 2022), categorize them into 10 broad domains with Llama-3-72B (Grattafiori et al., 2024), and use GPT-4o (Hurst et al., 2024) to segment each video into non-overlapping events from chronologically arranged 1 FPS frame sequences. A temporal coherence filter removes weak event chains before masked event supervision is created. VER contains 500K untrimmed videos, spans 18K hours, and provides 700K training instances with dense event captions and structured reasoning targets.

Under task-zero-shot evaluation, i.e., without fine-tuning on the target benchmarks, TEMPURA improves the Qwen2.5-VL-3B baseline by 6.1 mIoU on Charades-STA and 6.9 HIT@1 on QVHighlights. The pipeline is model-agnostic, and the gains hold across model families and scales: TEMPURA improves InternVL3-2B by 13.9 mIoU, Qwen2.5-VL-7B by 2.8 mIoU and 14.1 HIT@1, and InternVL3-8B by 10.3 mIoU and 23.7 HIT@1. When used as a pre-training stage before task-specific fine-tuning, the improvements grow further to 47.7 mIoU on Charades-STA and 75.1 HIT@1 on QVHighlights. These gains come without sacrificing broader capability: all four TEMPURA-trained models maintain their general video understanding performance on Video-MME (Fu et al., 2025). Ablations show that masked event prediction is most effective when it precedes dense captioning: on its own it acts as event-centric temporal pretraining rather than a grounding objective, and applying it first outperforms both dense captioning alone and the reversed order, supporting our design of event reasoning before event grounding. Our contributions are threefold:

- We develop TEMPURA, a two-stage training pipeline that first learns masked event reasoning and then learns dense event segmentation and captioning for grounded video understanding.
- We curate VER, a large-scale dataset of 500K videos and 18K hours with timestamp-aligned event captions and masked-event reasoning supervision across 10 common video categories.
- We show that the pipeline is model-agnostic and scalable: it improves Qwen2.5-VL and InternVL3 backbones from 2B to 8B on temporal grounding and highlight detection while preserving general video understanding.

2 Related Work

2.1 Video Large Multi-modal Models

Researchers have developed video LMMs for broad video understanding tasks (Zhao et al., 2024a;b). Many methods combine vision foundation models (Zhai et al., 2023; Radford et al., 2021) with LLMs (Grattafiori et al., 2024; Cai et al., 2024; Yang et al., 2024) and rely on token merging or compression to process longer videos (Song et al., 2024; Jin et al., 2024; Xu et al., 2024; Wu, 2024; Song et al., 2026; Chai et al., 2025). LLaVA-OneVision and LLaVA-Next-Interleave extend the LLaVA design to image and video settings with strong general performance (Liu et al., 2023; Li et al., 2025a; 2024a), while Video-LLaMA adds audio for richer multimodal understanding (Zhang et al., 2023; Cheng et al., 2024). Recent work shows that many of these models still struggle with temporal order and event localization (Ren et al., 2024; Huang et al., 2024; Guo et al., 2025b). We target this weakness with event-level supervision that preserves temporal structure instead of collapsing it.

2.2 LLM with Reasoning

LLMs now solve many multi-step problems through stronger reasoning-oriented training (Guo et al., 2025a). Parallel work in LMMs improves image reasoning with step-by-step supervision for math, sketch, design, localization and agentic setup (Zhang et al., 2025a; Shao et al., 2024; Peng et al., 2024; Vinker et al., 2025; Peng et al., 2025b; Doveh et al., 2025; Peng et al., 2025a; Li et al., 2024c; Peng et al., 2023; Wu et al., 2024; Peng et al., 2026). Methods such as LLaVA-CoT guide image reasoning through staged decomposition (Xu et al., 2025). Video reasoning remains less explored because a model must preserve temporal order and causal context across dynamic sequences. TEMPURA transfers structured reasoning supervision to this setting through masked event prediction.

2.3 Temporal Understanding with LMMs

Existing video LMMs improve timestamp grounding with adaptive compression, segment alignment, or refined temporal embeddings (Chen et al., 2024a; Huang et al., 2024; Qian et al., 2024; Guo et al., 2025b; Zeng et al., 2025; Ren et al., 2024). These methods focus on temporal localization or captioning, but they usually do not train a model to infer missing events or explain event-to-event causality. TEMPURA complements temporal grounding with masked event reasoning, which improves video coherence and fine-grained event understanding. Two recent works are closely related. MECD (Chen et al., 2024b) formulates multi-event causal discovery and infers explicit causal graphs among video events; in contrast, TEMPURA uses masked event prediction only as a pretraining objective for downstream grounding and does not attempt formal causal discovery. What Happens When (Ahn et al., 2026) shows that video LMMs often rely on language priors rather than visual-temporal evidence when reasoning about event order. TEMPURA mitigates this risk in its second stage by grounding every event description to non-overlapping timestamped segments, which requires alignment with the observed video content.

3 Method

To equip video LMMs with the ability to segment videos into meaningful events, establish temporal order, and infer relationships between events, we propose a two-stage training framework. The first stage, Masked Event Prediction, trains the model to infer missing events and reason about event dependencies from video context. The second stage, Video Segmentation and Dense Captioning, teaches the model to decompose videos into temporally grounded event sequences. Together, these stages build a structured understanding of video narratives that generalizes to downstream tasks such as temporal grounding and highlight detection.

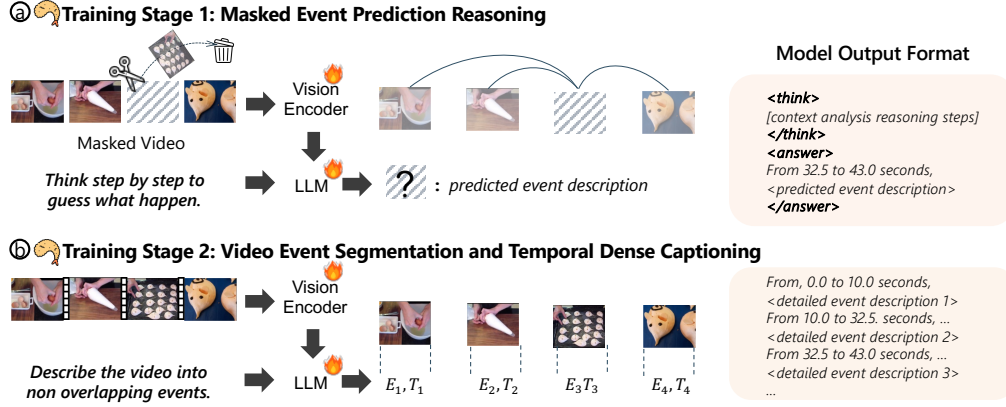


Figure 2: Overview of TEMPURA’s two-stage training pipeline. **(a) Masked Event Prediction Reasoning:** The model infers missing events from masked video context and produces a textual description with step-by-step causal explanations. **(b) Video Event Segmentation and Temporal Dense Captioning:** The model partitions an untrimmed video into non-overlapping events with start/end timestamps and detailed captions.

3.1 Masked Event Prediction

To strengthen video reasoning, we introduce **Masked Event Prediction**, a training stage that teaches the video LMM to infer a missing event from its surrounding context. Inspired by Fill-in-the-Middle (FIM) (Bavarian et al., 2022; Shen et al., 2023), we extend infilling from text to event sequences. Given dense event captions, we mask one event and ask a strong LLM to generate a plausible pseudo-event and a step-by-step explanation that links the missing event to the visible past and future events. Figure 2a shows this supervision format.

These pseudo-events and reasoning traces serve as supervised fine-tuning targets. The objective aligns video-based inference with language-based causal reasoning and encourages the model to recover logical event progression from video input alone.

3.2 Video Segmentation and Dense Captioning

In the second training stage, we teach the model to break down a video into non-overlapping events and describe each event in detail. As illustrated in Figure 2b, the video LMM learns to segment a video into non-overlapping events, each defined by its start and end timestamps, and generates detailed captions for each segment. Unlike Trace (Guo et al., 2025b), which utilizes extra encoders to model time and saliency scores, we train the model to directly ground all video segments using their enclosing timestamps, leveraging dense video captions from our VER dataset. This eliminates the need for additional parameters, making the video LMM more versatile while ensuring that it learns the structural and temporal progression of videos.

As summarized in Table 1, existing datasets lack large-scale timestamp-aligned dense event captions with full video coverage. To support TEMPURA training, we construct Video Event Reasoning (VER), a large-scale dataset of 500K untrimmed videos spanning 18K hours, with non-overlapping events, detailed descriptions, and temporal event reasoning data for masked event prediction. Unlike prior datasets with sparse or partial annotations, VER pairs full-coverage, timestamp-aligned dense captions with masked-event reasoning targets that are retained only when the surrounding event chain passes a temporal coherence check, yielding event-centric supervision.

Dataset Construction. Figure 3 presents our VER data pipeline. Starting from YT-1B (Zellers et al., 2022), we remove static videos following Farré et al. (2024) and categorize the remaining videos into 10 predefined categories using Llama-3-72B (Grattafiori

Table 1: Video Dataset Characteristics Comparison across mainstream benchmarks. Note that Charades contains overlapping temporal action annotations.

Dataset	Video Hours	Events per Video	Events per Minute	Coverage	Event Details	Temporal Reasoning
YouCook2 (Zhou et al., 2018)	175	7.7	1.5	Sparse	✓	×
Charades (Gao et al., 2017)	82.3	6.8	13.5	Sparse	✓	×
ActivityNet Captions (Krishna et al., 2017)	849	3.6	1.4	Sparse	×	×
FineVideo (Farré et al., 2024)	3,425	9.9	2.1	Dense	✓	×
ViTT (Huang et al., 2020)	541	7.1	1.8	Sparse	✓	×
Moment-10M (Qian et al., 2024)	7,260	22.5	3.3	Dense	✓	×
VER (Ours)	18,329	10.5	6.0	Dense	✓	✓

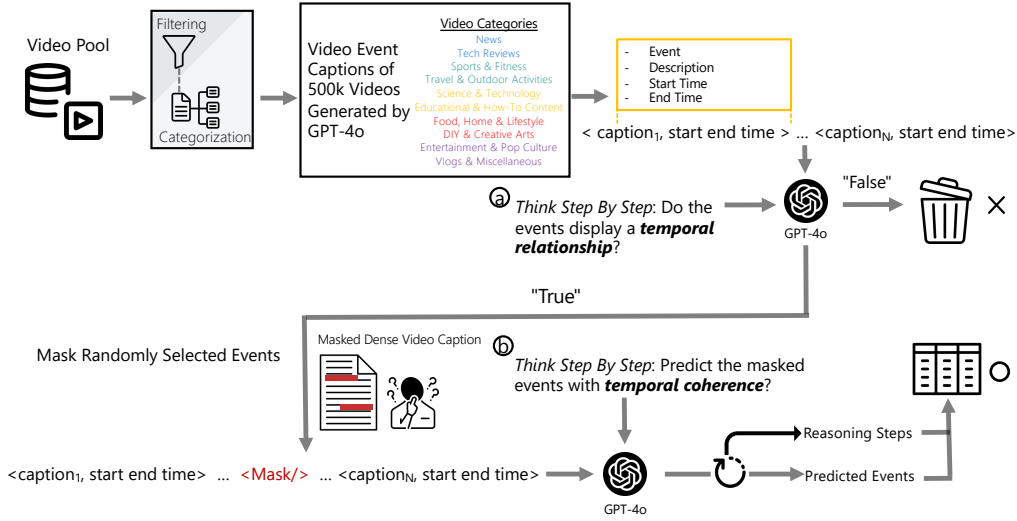


Figure 3: VER data pipeline. We filter and categorize a large video pool, generate dense event captions with GPT-4o, verify temporal coherence, and then create masked-event reasoning targets by masking valid event chains.

et al., 2024). To define event boundaries, we sample videos at 1 FPS, arrange frames into chronologically ordered frame sequence images, and apply GPT-4o (Hurst et al., 2024) to segment and caption each event. After filtering and alignment, we retain 500K videos with dense event captions, where each event includes a description and start/end timestamps (see Figure 4). To create training data for masked event prediction, we randomly mask an event from each dense caption and employ GPT-4o to predict the missing event within the masked time window. To ensure masked events are logically inferable, we filter out videos with uncorrelated event captions by prompting GPT-4o to assess causal relationships via step-by-step reasoning. During training, we align the video LMM’s reasoning with LLM event inference by fine-tuning on these structured reasoning processes, as shown in Figure 4.

4 Experiments

We fully fine-tuned the base models on the VER dataset following the two-stage TEMPURA training pipeline.

4.1 Video Temporal Understanding

We evaluate TEMPURA on two video temporal understanding benchmarks in a *task-zero-shot* setting, i.e., the model is not fine-tuned on the target benchmark. On Charades-STA (Gao

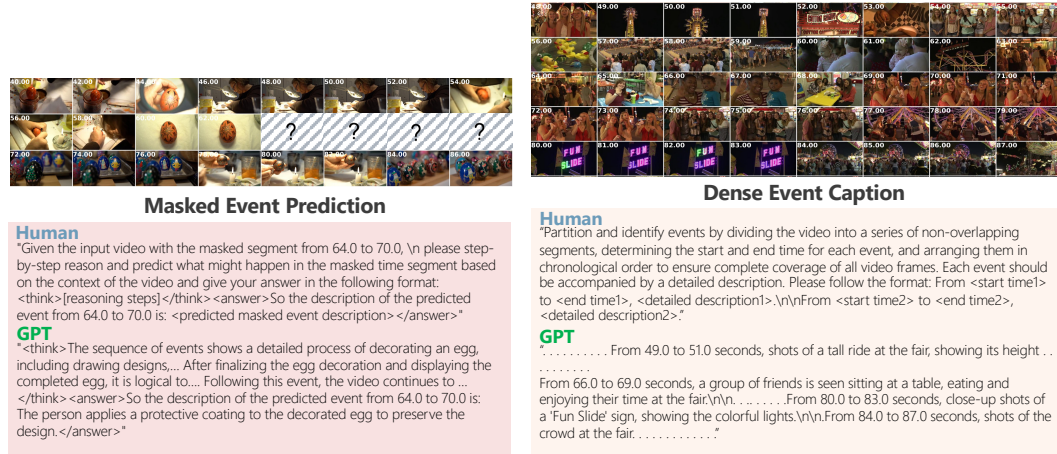


Figure 4: Structured Training Data for Masked Event Prediction and Dense Event Caption

et al., 2017), we report mean Intersection over Union (mIoU) and Recall@1 at IoU threshold τ (denoted $R@1@_t$). On QVHighlights (Lei et al., 2021), we report mean Average Precision (mAP) and HIT@1.

Table 2 shows that TEMPURA consistently improves every base model it is applied to. On Qwen2.5-VL-3B, TEMPURA improves Charades-STA by +6.1 mIoU and +5.0 $R@1@0.5$, and QVHighlights by +6.2 mAP and +6.9 HIT@1. On InternVL3-2B, the gains are even larger: +13.9 mIoU and +19.9 $R@1@0.5$ on Charades-STA, and +4.6 mAP and +3.6 HIT@1 on QVHighlights. The improvements persist at larger scales: TEMPURA improves Qwen2.5-VL-7B by +2.8 mIoU on Charades-STA and by +8.1 mAP and +14.1 HIT@1 on QVHighlights, and improves InternVL3-8B by +10.3 mIoU, +10.3 mAP, and +23.7 HIT@1. These gains are notable because TEMPURA is evaluated task-zero-shot and uses no target-task fine-tuning, extra temporal heads, or video-specific grounding modules (Li et al., 2024d; Huang et al., 2024; Ren et al., 2024; Wang et al., 2024; Guo et al., 2025b). We attribute the improvements to event-level supervision that teaches the model to segment videos, infer event transitions, and ground detailed captions in time. TEMPURA also strengthens task-specific fine-tuning: when both models are further fine-tuned with TimeChat’s (Ren et al., 2024) instruction-following data, TEMPURA-Qwen2.5-VL-3B (FT) outperforms the fine-tuned Qwen2.5-VL-3B by +7.4 mIoU and +11.8 $R@1@0.5$ on Charades-STA, and +23.1 mAP and +30.7 HIT@1 on QVHighlights. Together, the consistent improvements across the Qwen2.5-VL and InternVL3 families from 2B to 8B scales confirm that the benefits stem from the event-centric training signal rather than a specific architecture.

General Video Understanding. To test whether event-centric training narrows broader capability, we evaluate all backbones on Video-MME (Fu et al., 2025). As shown in Table 3, TEMPURA maintains overall accuracy on every base model (+0.22 to +1.07), indicating that the gains in temporal grounding and event localization are achieved while preserving general video understanding.

4.2 Ablation Study

To study the effectiveness of each component in TEMPURA, we split our ablation analysis into three parts and report mIoU and $R@1$ on Charades-STA. All ablations use Qwen2.5-VL-3B as the base model.

TEMPURA Training Stage. Table 4 shows that masked event prediction is most effective when it serves as the first stage before dense captioning. S1 alone is not a grounding objective: without timestamp-aligned supervision it reaches only 25.6 mIoU, functioning instead as event-centric temporal pretraining that teaches the model to infer missing events and model event continuity. The reverse order, S2 $\rightarrow$ S1, underperforms because the model

Table 2: **Video Temporal Grounding** on Charades-STA and **Highlight Detection** on QVHighlights. Top: prior methods fine-tuned on the benchmark training sets (FT). Middle: prior video LMMs evaluated without benchmark fine-tuning. Bottom: same-backbone pairs comparing each base model with its TEMPURA-trained counterpart under the task-zero-shot setting, i.e., without fine-tuning on the target benchmarks; in the last pair, FT denotes additional fine-tuning on the benchmark training sets. A dash indicates that the result was not reported under the corresponding benchmark, metric, and evaluation setting.

Method	Charades-STA				QVHighlights	
	mIoU	R@1 (IoU=0.3)	R@1 (IoU=0.5)	R@1 (IoU=0.7)	mAP	HIT@1
QD-DETR (FT) (Moon et al., 2023)	-	-	57.3	32.6	38.9	64.2
UnLoc-L (FT) (Yan et al., 2023)	-	-	60.8	38.4	-	-
HawkEye (FT) (Wang et al., 2024)	49.3	72.5	58.3	28.8	-	-
TimeChat (FT) (Ren et al., 2024)	-	-	46.7	23.7	21.7	37.9
VideoChat-T (FT) (Zeng et al., 2025)	-	79.4	67.1	43.0	27.0	55.3
MovieChat (Song et al., 2024)	-	8.8	2.9	1.3	11.7	16.1
GroundingGPT (Li et al., 2024d)	-	-	29.6	11.9	-	-
VTimeLLM (Huang et al., 2024)	31.2	51.0	27.5	11.4	-	-
HawkEye (Wang et al., 2024)	33.7	50.6	31.4	14.5	-	-
TimeChat (Ren et al., 2024)	-	-	32.2	13.4	14.5	23.9
TRACE (Guo et al., 2025b)	38.7	58.6	40.3	19.4	26.8	42.7
ChatVTG (Qu et al., 2024)	34.9	52.7	33.0	15.9	-	-
VideoChat2 (Li et al., 2024b)	-	9.6	3.4	1.4	13.4	18.6
Momentor (Qian et al., 2024)	28.5	42.6	26.6	11.6	7.6	-
Grounded-VideoLLM (Wang et al., 2025)	36.8	54.2	36.4	19.7	-	-
Qwen2.5-VL-3B (Bai et al., 2025)	33.1	52.4	34.3	12.5	42.1	44.8
TEMPURA-Qwen2.5-VL-3B	39.2 (+6.1)	63.9 (+11.5)	39.3 (+5.0)	15.0 (+2.5)	48.3 (+6.2)	51.7 (+6.9)
InternVL3-2B (Zhu et al., 2025)	34.3	56.5	29.7	10.9	24.7	21.2
TEMPURA-InternVL3-2B	48.2 (+13.9)	81.8 (+25.3)	49.6 (+19.9)	17.9 (+7.0)	29.3 (+4.6)	24.8 (+3.6)
Qwen2.5-VL-7B (Bai et al., 2025)	42.7	68.5	43.7	17.3	42.2	46.2
TEMPURA-Qwen2.5-VL-7B	45.5 (+2.8)	74.4 (+5.9)	47.9 (+4.2)	18.8 (+1.5)	50.3 (+8.1)	60.3 (+14.1)
InternVL3-8B (Zhu et al., 2025)	44.6	75.6	42.5	14.4	43.5	44.5
TEMPURA-InternVL3-8B	54.9 (+10.3)	91.3 (+15.7)	58.7 (+16.2)	23.7 (+9.3)	53.8 (+10.3)	68.2 (+23.7)
Qwen2.5-VL-3B (FT) (Bai et al., 2025)	40.3	59.0	43.1	24.7	34.7	44.4
TEMPURA-Qwen2.5-VL-3B (FT)	47.7 (+7.4)	65.9 (+6.9)	54.9 (+11.8)	36.3 (+11.6)	57.8 (+23.1)	75.1 (+30.7)

Table 3: Video-MME (Fu et al., 2025) overall accuracy. TEMPURA maintains general video understanding across all base models.

Backbone	Baseline	TEMPURA	Δ
Qwen2.5-VL-3B	54.44	55.07	+0.63
Qwen2.5-VL-7B	57.22	57.44	+0.22
InternVL3-2B	48.52	49.59	+1.07
InternVL3-8B	55.19	55.41	+0.22

is not first taught to reason about missing events and event transitions before it learns to localize segments, although we still observe that it extracts facts around the masked time windows and generates longer reasoning steps to predict plausible infills. In contrast, S1 $\rightarrow$ S2 outperforms both S2-only and the reversed curriculum, which supports our design of event reasoning before event grounding.

Table 4: TEMPURA training stages. **S1**: masked event prediction. **S2**: event segmentation and temporal captioning.

Training Stages	mIoU	R@1 (IoU=0.3)	R@1 (IoU=0.5)
S1	25.6	33.1	16.7
S2	38.4	59.1	32.8
S2 $\rightarrow$ S1	34.0	55.6	32.6
S1 $\rightarrow$ S2	39.2	63.9	39.3

Table 5: Dynamic scene and relevant segment filtering. **D.S.**: dynamic scene filtering. **T.R.**: relevant segment filtering.

Data Filtering	mIoU	R@1 (IoU=0.3)	R@1 (IoU=0.5)
No D.S.	34.8	51.3	27.4
D.S.	38.7 (+3.9)	63.7 (+12.4)	38.3 (+10.9)
No T.R.	33.0	47.3	28.3
T.R.	37.5 (+4.5)	57.1 (+9.8)	34.8 (+6.5)

Table 6: Performance on Charades-STA under different training frame sampling rates. Inference is fixed at 0.5 FPS for all settings.

Training FPS	mIoU	R@1 (IoU=0.5)	R@1 (IoU=0.7)
0.25	37.3	36.8	14.5
0.5	39.5	38.8	14.3
1.0	38.9	38.4	14.2

Dynamic Scene and Relevant Event Filtering. Table 5 shows that filtering static scenes improves dense captioning supervision, while filtering temporally irrelevant event chains improves masked-event reasoning supervision. Both data filters matter because TEMPURA learns from event progression. Static or weakly related segments dilute the temporal signal and reduce grounding quality.

Training Frame Sampling Rate. While VER event boundaries are annotated using 1 FPS frame sampling with GPT-4o, the frame sampling rate used during model training is a separate choice. We investigate this by training TEMPURA at 0.25, 0.5, and 1.0 FPS while holding inference fixed at 0.5 FPS. As shown in Table 6, TEMPURA performs robustly across all three settings, with mIoU ranging narrowly from 37.3 to 39.5.

4.3 Discussion

The ablations support the central design choice in TEMPURA: event reasoning should come before dense temporal grounding. Stage 1 does not simply expose the model to more captions. It changes the supervision signal. When the model predicts a masked event from visible past and future context, it must represent event transitions, preconditions, and likely consequences. This objective differs from standard dense captioning, where the model can often match local appearance without modeling why one event should follow another. The stage-order ablation in Table 4 reflects this difference. S1 $\rightarrow$ S2 gives the strongest performance because stage 1 builds an event-level prior over temporal structure, and stage 2 then grounds that prior with explicit boundaries and captions. The reverse order exposes the model to dense event descriptions first, but it does not encourage the model to reason over missing structure before it learns to localize segments.

The downstream gains also align with how TEMPURA decomposes videos. Charades-STA rewards precise alignment between a text query and a short temporal window, while QVHighlights requires the model to rank multiple relevant windows instead of a single best span. Both tasks benefit when the model already represents the video as an ordered event sequence rather than a flat stream of frames. Figure 5 and videos on the project page show

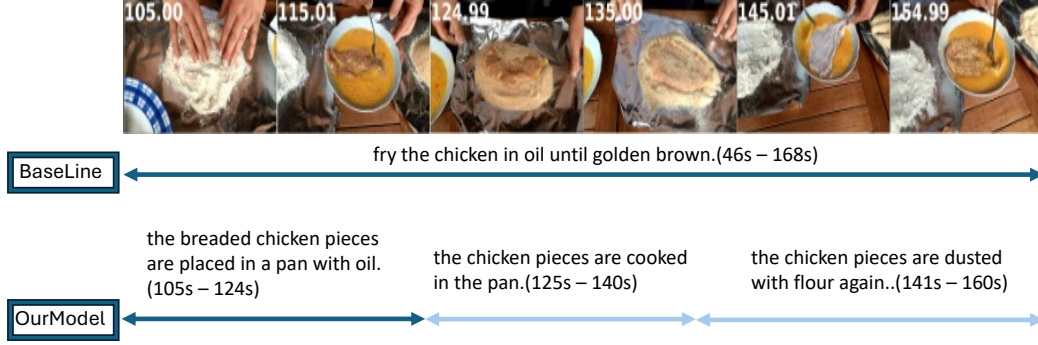


Figure 5: TEMPURA produces finer-grained event partitions than its base model, Qwen2.5-VL, enabling more precise temporally grounded captions.

that TEMPURA infers more coherent segments than Qwen2.5-VL, which leads to better grounded captions for small transitions.

In Table 2 fine-tuning (FT) experiments, when TEMPURA is used as a pre-training stage before task-specific fine-tuning, the gains over the baseline grow substantially (+7.4 mIoU on Charades-STA, +30.7 HIT@1 on QVHighlights). This suggests that the event-centric representation learned during TEMPURA’s two stages provides a strong foundation that task-specific supervision can build upon. TEMPURA training stages also yields large improvements on InternVL3-2B (+13.9 mIoU) confirms that the benefits are not tied to a specific architecture but stem from the training signal itself.

5 Limitations and Future Work

TEMPURA still leaves several open problems. First, event boundaries are inherently ambiguous in videos that contain repetition, gradual state changes, or overlapping sub-actions. A single “correct” segmentation may not exist. TEMPURA learns one segmentation that is useful for training, but downstream users may prefer coarser or finer partitions depending on the task. This issue is especially visible in cooking, sports, and vlog-style videos, where small hand motions, object state changes, and camera movement all compete as possible event boundaries. A promising direction is to model boundary uncertainty explicitly rather than forcing one deterministic event chain for every clip.

Second, VER trades annotation scale for annotation purity. GPT-4o provides dense event captions, temporal coherence checks, and masked-event supervision, which makes the dataset large enough to train modern video LMMs. The same design also introduces noise: some captions are too coarse, some boundaries are slightly shifted, and some inferred masked events are only one of several plausible alternatives. Better filtering helps, but it cannot fully solve the problem. A natural next step is to combine automatic annotation with targeted human verification on difficult clips, or to use confidence-weighted training objectives that down-weight uncertain event chains. Another limitation is modality coverage. TEMPURA uses visual supervision only. Audio cues often disambiguate event transitions in instructional, sports, and conversational videos, so multi-modal event supervision may improve the representation substantially.

Third, our evaluation focuses on temporal grounding and highlight detection, which test whether the learned representation transfers to downstream temporal tasks. These benchmarks do not fully measure open-ended reasoning over long videos. Future work should test whether event-centric supervision improves multi-hop video QA, video planning, or interactive retrieval, where a model must compose several events before it answers. The same applies to agents that execute long-horizon tasks alongside people, since an agent must decide when to step in (Yang et al., 2026; Peng et al., 2025b). It would also be useful to study how TEMPURA interacts with external memory, retrieval augmentation, or planner-style decoders for videos that are much longer than the current training distribution. We view

TEMPURA as one part of a broader recipe: learn event continuity, ground events densely in time, and then build higher-level reasoning modules on top of that representation.

6 Conclusions

We present TEMPURA, a two-stage training framework that improves video temporal understanding through masked event reasoning and dense event grounding. We also introduce VER, a large-scale dataset that provides both dense event captions and structured reasoning targets. TEMPURA substantially improves base video LMMs on task-zero-shot temporal grounding and highlight detection, yielding consistent gains across the Qwen2.5-VL and InternVL3 families from 2B to 8B scales, and ablations show that reasoning-first training is important for fine-grained video understanding. The model-agnostic pipeline also serves as an effective pre-training stage that amplifies task-specific fine-tuning gains on both Charades-STA and QVHighlights. More broadly, our results suggest that video LMMs benefit from training signals that model events as structured temporal units instead of treating videos as flat frame sequences.

Our findings also point to a practical recipe for future video LMM training. A model first needs supervision that captures continuity across events, because causal ordering is difficult to recover from local appearance alone. Once the model acquires that prior, dense timestamped captions can ground the representation and make it useful for downstream localization and retrieval tasks. The strong fine-tuning results confirm that this event-centric prior transfers well: task-specific supervision builds on the structured representation rather than replacing it. We expect this recipe to complement stronger memory systems, retrieval modules, and planning-oriented decoders rather than compete with them. We hope this perspective motivates future work on scalable event supervision for multimodal reasoning over long-form video.

At a systems level, TEMPURA suggests that better temporal supervision can substitute for some architectural complexity. Several prior approaches improve long-video understanding with specialized temporal heads, additional retrieval components, or benchmark-specific decoding rules. TEMPURA instead improves the base model through an event-centric training signal that is simple to apply at fine-tuning time. This design does not remove the value of specialized components, but it changes where complexity is spent. Rather than adding more inference-time machinery, one can invest in higher-quality event annotations, stronger causal filtering, and denser timestamped supervision. We believe this trade-off is attractive for multimodal research because it scales across tasks and across architectures: once a model learns to represent video as a chain of grounded events, the same representation can support grounding, highlighting, captioning, retrieval, and more open-ended reasoning.

References

- Daechul Ahn, Yura Choi, Hyeonbeom Choi, Seongwon Cho, San Kim, and Jonghyun Choi. What happens when: Learning temporal orders of events in videos. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 2786–2796, 2026. arXiv:2512.08979.
- Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao 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, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-VL technical report, 2025. URL <https://arxiv.org/abs/2502.13923>.
- Mohammad Bavarian, Heewoo Jun, Nikolas Tezak, John Schulman, Christine McLeavey, Jerry Tworek, and Mark Chen. Efficient training of language models to fill in the middle, 2022. URL <https://arxiv.org/abs/2207.14255>.
- Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, Xiaoyi Dong, et al. InternLM2 technical report, 2024.
- Wenhao Chai, Enxin Song, Yilun Du, Chenlin Meng, Vashisht Madhavan, Omer Bar-Tal, Jenq-Neng Hwang, Saining Xie, and Christopher D Manning. AuroraCap: Efficient, performant video detailed captioning and a new benchmark. In *International Conference on Learning Representations (ICLR)*, 2025. arXiv:2410.03051.
- Shimin Chen, Xiaohan Lan, Yitian Yuan, Zequn Jie, and Lin Ma. TimeMarker: A versatile video-LLM for long and short video understanding with superior temporal localization ability, 2024a. URL <https://arxiv.org/abs/2411.18211>.
- Tieyuan Chen, Huabin Liu, Tianyao He, Yihang Chen, Chaofan Gan, Xiao Ma, Cheng Zhong, Yang Zhang, Yingxue Wang, Hui Lin, and Weiyao Lin. MECD: Unlocking multi-event causal discovery in video reasoning. In *Advances in Neural Information Processing Systems*, 2024b. arXiv:2409.17647.
- Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, et al. VideoLLaMA 2: Advancing spatial-temporal modeling and audio understanding in Video-LLMs. *arXiv preprint arXiv:2406.07476*, 2024.
- Sivan Doveh, Nimrod Shabtay, Eli Schwartz, Hilde Kuehne, Raja Giryes, Rogerio Feris, Leonid Karlinsky, James Glass, Assaf Arbelle, Shimon Ullman, et al. Teaching vlms to localize specific objects from in-context examples. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9572–9582. IEEE, 2025.
- Miquel Farré, Andi Marafioti, Lewis Tunstall, Leandro Von Werra, and Thomas Wolf. FineVideo. <https://huggingface.co/datasets/HuggingFaceFV/finevideo>, 2024.
- Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-MME: The first-ever comprehensive evaluation benchmark of multi-modal LLMs in video analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24108–24118, 2025. arXiv:2405.21075.
- Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. TALL: Temporal activity localization via language query. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 5277–5285, 2017. doi: 10.1109/ICCV.2017.563.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.
- Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645:633–638, 2025a.

- Yongxin Guo, Jingyu Liu, Mingda Li, Xiaoying Tang, Qingbin Liu, and Xi Chen. TRACE: Temporal grounding video LLM via causal event modeling. In *International Conference on Learning Representations (ICLR)*, 2025b. arXiv:2410.05643.
- Bin Huang, Xin Wang, Hong Chen, Zihan Song, and Wenwu Zhu. VTimeLLM: Empower LLM to grasp video moments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14271–14280, 2024.
- Gabriel Huang, Bo Pang, Zhenhai Zhu, Clara Rivera, and Radu Soricut. Multimodal pretraining for dense video captioning. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pp. 470–490, 2020.
- Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. GPT-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.
- Peng Jin, Ryuichi Takanobu, Wancai Zhang, Xiaochun Cao, and Li Yuan. Chat-UniVi: Unified visual representation empowers large language models with image and video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13700–13710, 2024.
- Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 706–715, 2017.
- Jie Lei, Tamara L. Berg, and Mohit Bansal. QVHighlights: Detecting moments and highlights in videos via natural language queries. In *Proceedings of the 35th International Conference on Neural Information Processing Systems, NIPS '21*, Red Hook, NY, USA, 2021. Curran Associates Inc. ISBN 9781713845393.
- Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. LLaVA-OneVision: Easy visual task transfer. *Transactions on Machine Learning Research*, 2025a. arXiv:2408.03326.
- Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. LLaVA-NeXT-Interleave: Tackling multi-image, video, and 3D in large multimodal models. *arXiv preprint arXiv:2407.07895*, 2024a.
- Kunchang Li, Yali Wang, Yanan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. MVBench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22195–22206, 2024b.
- Ling Li, Yu Ye, Yao Zhou, Bingchuan Jiang, and Wei Zeng. Georeasoner: Geo-localization with reasoning in street views using a large vision-language model. *arXiv preprint arXiv:2406.18572*, 2024c.
- Rui Li, Xiaohan Wang, Yuhui Zhang, Orr Zohar, Zeyu Wang, and Serena Yeung-Levy. Temporal preference optimization for long-form video understanding. *arXiv preprint arXiv:2501.13919*, 2025b.
- Zhaowei Li, Qi Xu, Dong Zhang, Hang Song, Yiqing Cai, Qi Qi, Ran Zhou, Junting Pan, Zefeng Li, Van Tu Vu, et al. GroundingGPT: Language enhanced multi-modal grounding model. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6657–6678, 2024d.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in Neural Information Processing Systems*, 36:34892–34916, 2023.
- Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-ChatGPT: Towards detailed video understanding via large vision and language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12585–12602, 2024.

- WonJun Moon, Sangeek Hyun, SangUk Park, Dongchan Park, and Jae-Pil Heo. Query-dependent video representation for moment retrieval and highlight detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23023–23033, 2023.
- Yi-Hao Peng, Peggy Chi, Anjuli Kannan, Meredith Ringel Morris, and Irfan Essa. Slide gestalt: Automatic structure extraction in slide decks for non-visual access. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pp. 1–14, 2023.
- Yi-Hao Peng, Faria Huq, Yue Jiang, Jason Wu, Xin Yue Li, Jeffrey P Bigham, and Amy Pavel. Dreamstruct: Understanding slides and user interfaces via synthetic data generation. In *European Conference on Computer Vision*, pp. 466–485. Springer, 2024.
- Yi-Hao Peng, Jeffrey P Bigham, and Jason Wu. Designpref: Capturing personal preferences in visual design generation. *arXiv preprint arXiv:2511.20513*, 2025a.
- Yi-Hao Peng, Dingzeyu Li, Jeffrey P. Bigham, and Amy Pavel. Morae: Proactively pausing UI agents for user choices. In *Proceedings of the 38th Annual ACM Symposium on User Interface Software and Technology (UIST)*, 2025b.
- Yi-Hao Peng, Samarth Das, Jeffrey P. Bigham, and Jason Wu. Efficient personalization of generative user interfaces. *arXiv preprint arXiv:2604.09876*, 2026.
- Long Qian, Juncheng Li, Yu Wu, Yaobo Ye, Hao Fei, Tat-Seng Chua, Yueting Zhuang, and Siliang Tang. Momentor: Advancing video large language model with fine-grained temporal reasoning. In *Proceedings of the 41st International Conference on Machine Learning, ICML’24*. JMLR.org, 2024.
- Mengxue Qu, Xiaodong Chen, Wu Liu, Alicia Li, and Yao Zhao. ChatVTG: Video temporal grounding via chat with video dialogue large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 1847–1856, 2024.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pp. 8748–8763. PMLR, 2021.
- Shuhuai Ren, Linli Yao, Shicheng Li, Xu Sun, and Lu Hou. TimeChat: A time-sensitive multimodal large language model for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14313–14323, 2024.
- Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. Visual CoT: Advancing multi-modal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, 2024. arXiv:2403.16999.
- Tianxiao Shen, Hao Peng, Ruoqi Shen, Yao Fu, Zaid Harchaoui, and Yejin Choi. FiLM: Fill-in language models for any-order generation, 2023. URL <https://arxiv.org/abs/2310.09930>.
- Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Haozhe Chi, Xun Guo, Tian Ye, Yanting Zhang, et al. MovieChat: From dense token to sparse memory for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18221–18232, 2024.
- Enxin Song, Wenhao Chai, Tian Ye, Jenq-Neng Hwang, Xi Li, and Gaoang Wang. MovieChat+: Question-aware sparse memory for long video question answering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(1):374–389, 2026.
- Yael Vinker, Tamar Rott Shaham, Kristine Zheng, Alex Zhao, Judith E Fan, and Antonio Torralba. SketchAgent: Language-driven sequential sketch generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. arXiv:2411.17673.

- Haibo Wang, Zhiyang Xu, Yu Cheng, Shizhe Diao, Yufan Zhou, Yixin Cao, Qifan Wang, Weifeng Ge, and Lifu Huang. Grounded-VideoLLM: Sharpening fine-grained temporal grounding in video large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, 2025. arXiv:2410.03290.
- Yueqian Wang, Xiaojun Meng, Jianxin Liang, Yuxuan Wang, Qun Liu, and Dongyan Zhao. HawkEye: Training video-text LLMs for grounding text in videos. *arXiv preprint arXiv:2403.10228*, 2024.
- Jason Wu, Yi-Hao Peng, Xin Yue Amanda Li, Amanda Swearngin, Jeffrey P. Bigham, and Jeffrey Nichols. UIClip: A data-driven model for assessing user interface design. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology (UIST '24)*. Association for Computing Machinery, 2024. doi: 10.1145/3654777.3676408.
- Wenhao Wu. FreeVA: Offline MLLM as training-free video assistant. *arXiv preprint arXiv:2405.07798*, 2024.
- Guowei Xu, Peng Jin, Ziang Wu, Hao Li, Yibing Song, Lichao Sun, and Li Yuan. LLaVA-CoT: Let vision language models reason step-by-step. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025. arXiv:2411.10440.
- Mingze Xu, Mingfei Gao, Zhe Gan, Hong-You Chen, Zhengfeng Lai, Haiming Gang, Kai Kang, and Afshin Dehghan. SlowFast-LLaVA: A strong training-free baseline for video large language models. *arXiv preprint arXiv:2407.15841*, 2024.
- Shen Yan, Xuehan Xiong, Arsha Nagrani, Anurag Arnab, Zhonghao Wang, Weina Ge, David Ross, and Cordelia Schmid. UnLoc: A unified framework for video localization tasks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 13623–13633, 2023.
- An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. Qwen2.5 technical report, 2024. URL <https://arxiv.org/abs/2412.15115>.
- Saelyne Yang, Jaesang Yu, Yi-Hao Peng, Kevin Qinghong Lin, Jae Won Cho, Yale Song, and Juho Kim. GUIDE: A benchmark for understanding and assisting users in open-ended GUI tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026. arXiv:2603.25864.
- Rowan Zellers, Jiasen Lu, Ximing Lu, Youngjae Yu, Yanpeng Zhao, Mohammadreza Salehi, Aditya Kusupati, Jack Hessel, Ali Farhadi, and Yejin Choi. MERLOT reserve: Neural script knowledge through vision and language and sound. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16375–16387, 2022.
- Xiangyu Zeng, Kunchang Li, Chenting Wang, Xinhao Li, Tianxiang Jiang, Ziang Yan, Songze Li, Yansong Shi, Zhengrong Yue, Yi Wang, Yali Wang, Yu Qiao, and Limin Wang. TimeSuite: Improving MLLMs for long video understanding via grounded tuning. In *International Conference on Learning Representations (ICLR)*, 2025. arXiv:2410.19702.
- Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11975–11986, 2023.
- Hang Zhang, Xin Li, and Lidong Bing. Video-LLaMA: An instruction-tuned audio-visual language model for video understanding. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 543–553, 2023.
- Renrui Zhang, Xinyu Wei, Dongzhi Jiang, Ziyu Guo, Shicheng Li, Yichi Zhang, Chengzhuo Tong, Jiaming Liu, Aojun Zhou, Bin Wei, Shanghang Zhang, Peng Gao, Chunyuan Li, and Hongsheng Li. MAVIS: Mathematical visual instruction tuning with an automatic data engine. In *International Conference on Learning Representations (ICLR)*, 2025a. arXiv:2407.08739.

- Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. LLaVA-Video: Video instruction tuning with synthetic data. *Transactions on Machine Learning Research*, 2025b. arXiv:2410.02713.
- Zhonghan Zhao, Wenhao Chai, Xuan Wang, Boyi Li, Shengyu Hao, Shidong Cao, Tian Ye, and Gaoang Wang. See and think: Embodied agent in virtual environment. In *European Conference on Computer Vision*, pp. 187–204. Springer, 2024a.
- Zhonghan Zhao, Kewei Chen, Dongxu Guo, Wenhao Chai, Tian Ye, Yanting Zhang, and Gaoang Wang. Hierarchical auto-organizing system for open-ended multi-agent navigation. *arXiv preprint arXiv:2403.08282*, 2024b.
- Luowei Zhou, Chenliang Xu, and Jason J Corso. Towards automatic learning of procedures from web instructional videos. In *AAAI Conference on Artificial Intelligence*, pp. 7590–7598, 2018. URL <https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/17344>.
- Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Yuchen Duan, Hao Tian, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Yue Cao, Yangzhou Liu, Weijia Wu, Wenhao Wang, Lewei Lu, Zhenjiang Jin, Chaoya Jiang, Xizhou Zhu, Yu Qiao, Dahua Lin, and Jifeng Dai. InternVL3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

A Appendix

The Appendix is structured as follows:

- VER data creation pipeline and statistics in Section B.
- Qualitative analysis in Section C.

B VER Data Creation Pipeline and Statistics

After data filtering, we uniformly sampled video frames and arranged them into frame sequence images. The example input to GPT-4o is shown on the right of Figure 6. We overlaid the timestamp in each image, and combined multiple frame sequence images to obtain a grid-formatted composite image as the input. These inputs were first temporally segmented by GPT-4o, and we consider these segments as events. Based on the generated time segments, GPT-4o then generated the descriptions of these events separately. Next, we used these temporally aligned event descriptions to construct our masked event prediction data. To create this dataset, we first need to filter out data with weak event correlations. The bottom half of Figure 6 shows an example of the data pipeline described in Figure 3. By inputting the event information into the model in text form, we have it determine whether each correlation is logically valid. After filtering, we obtained 200K masked-event reasoning instances from the 500K dense video captions.

Our dataset contains videos across 10 domains like travel, DIY, tech reviews, etc. (Figure 7). The average duration and the length of the caption also varied between different domains (Figure 8). Figure 9 presents the percentage of videos with temporal relevance in each category.

C Qualitative Analysis

Figure 10 and 11 compare the performance of Qwen2.5-VL-3B (Bai et al., 2025) (our baseline), Grounded-Video-LLM-Meta-Llama-3-8B-Instruct (Wang et al., 2025), Grounded-Video-LLM-Phi (Wang et al., 2025), and TEMPURA (our proposed model) on long video temporal grounding tasks. The red text highlights errors in timestamp predictions when other models segment videos into fine-grained events and identify their start and end times. While other models often struggle, especially toward the end of long videos, TEMPURA consistently segments events accurately and assigns precise timestamps. For instance, the green text shows that TEMPURA correctly identifies a person filling and wrapping spring rolls from 161.00 to 183.00 seconds, followed by placing them into oil for frying from 185.00 to 205.00 seconds. Additionally, TEMPURA has better performance in producing more fine-grained event captions, as shown by the larger number of event captions produced by our model.

Figure 12 compares the performance of Qwen2.5-VL-3B (our baseline), Grounded-Video-LLM-Meta-Llama-3-8B-Instruct (Wang et al., 2025), Grounded-Video-LLM-Phi (Wang et al., 2025), and TEMPURA (our proposed model) on fine-grained video segmentation. The red highlights indicate errors in timestamp predictions and the failure of other models to produce detailed event captions, even in shorter videos. In contrast, TEMPURA demonstrates better performance, as indicated by the green text, by producing more accurate timestamps, fine-grained events, and descriptive event captions.

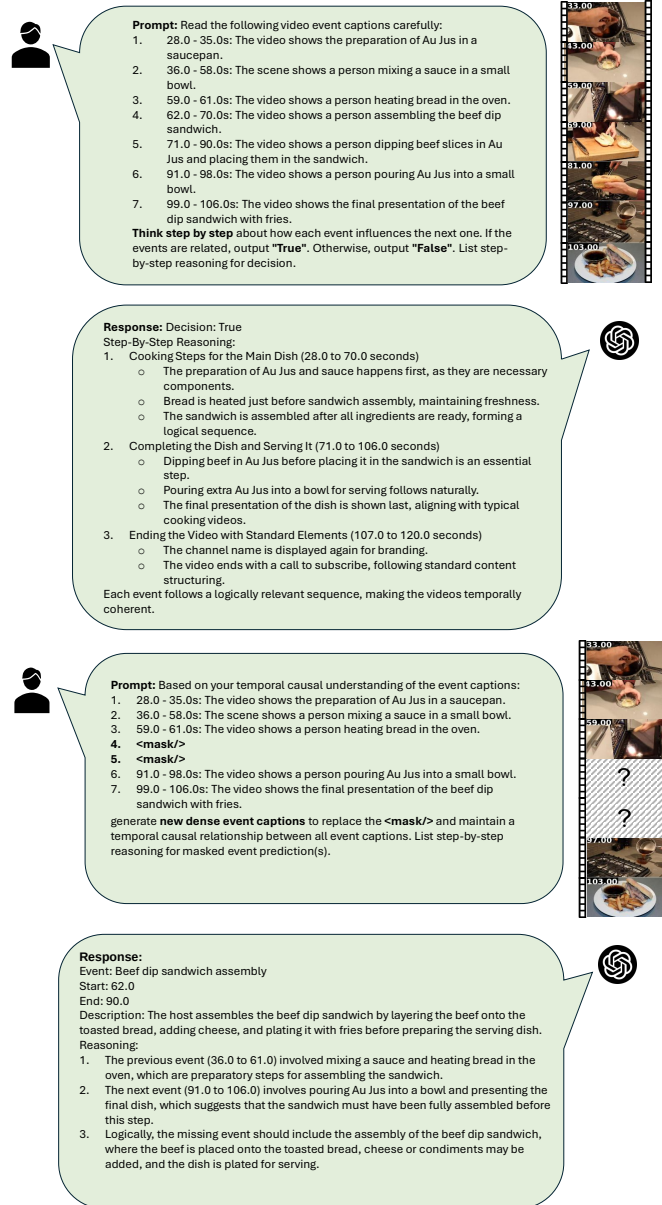


Figure 6: Our mask event prediction data example and generation process.

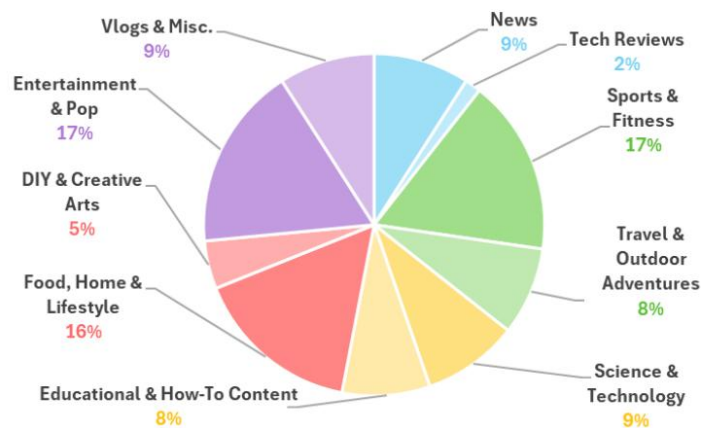


Figure 7: Video Class Distribution

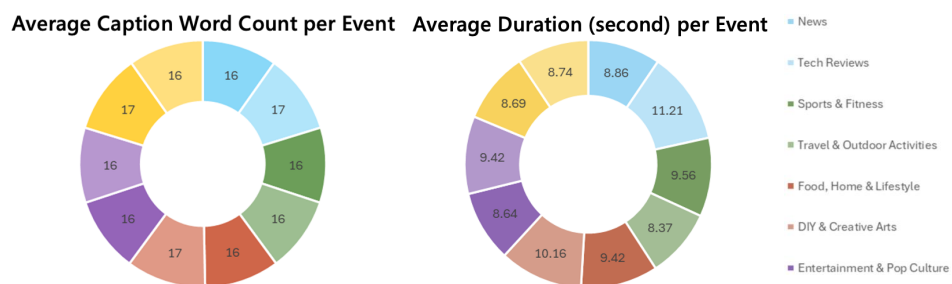


Figure 8: Event Caption Distribution

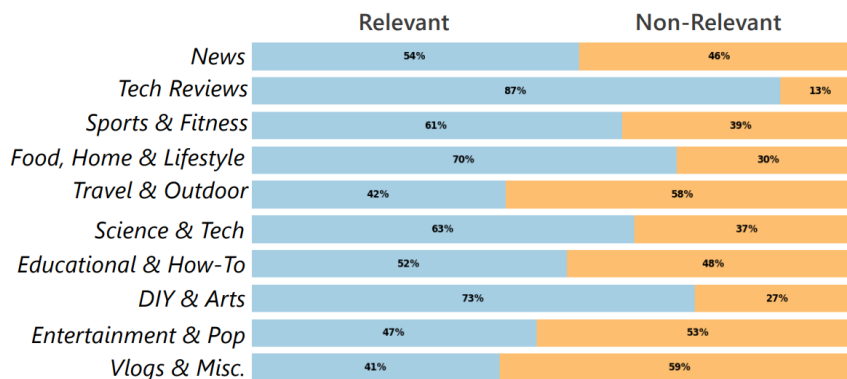


Figure 9: Video Frame Temporal Relevance Distribution

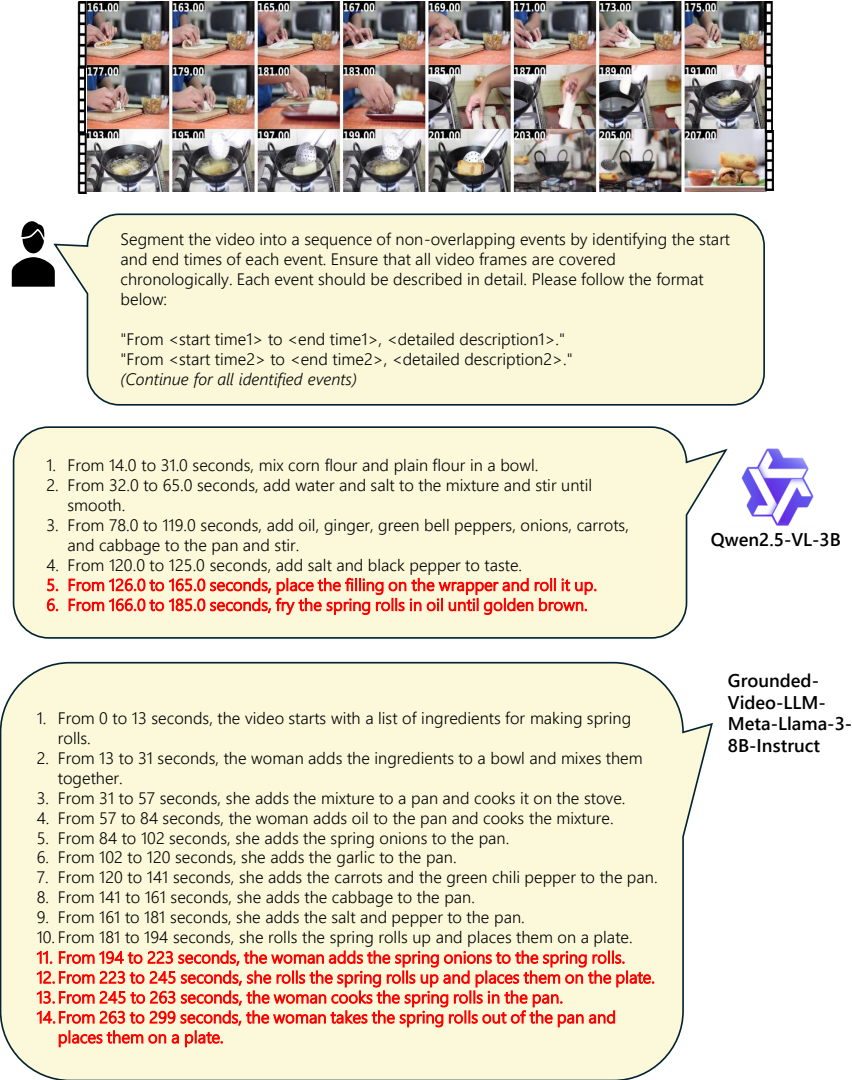


Figure 10: Comparison of **long video temporal grounding responses** on a cooking tutorial video, generated by baseline models: Qwen2.5-VL-3B (Bai et al., 2025) and Grounded-Video-LLM-Meta-Llama-3-8B-Instruct (Wang et al., 2025). Each of the baseline models is prompted to segment videos and describe each of the video segments in detail with correct start and end times. The text highlighted in red indicates incorrect determination of start and end times for frame descriptions.

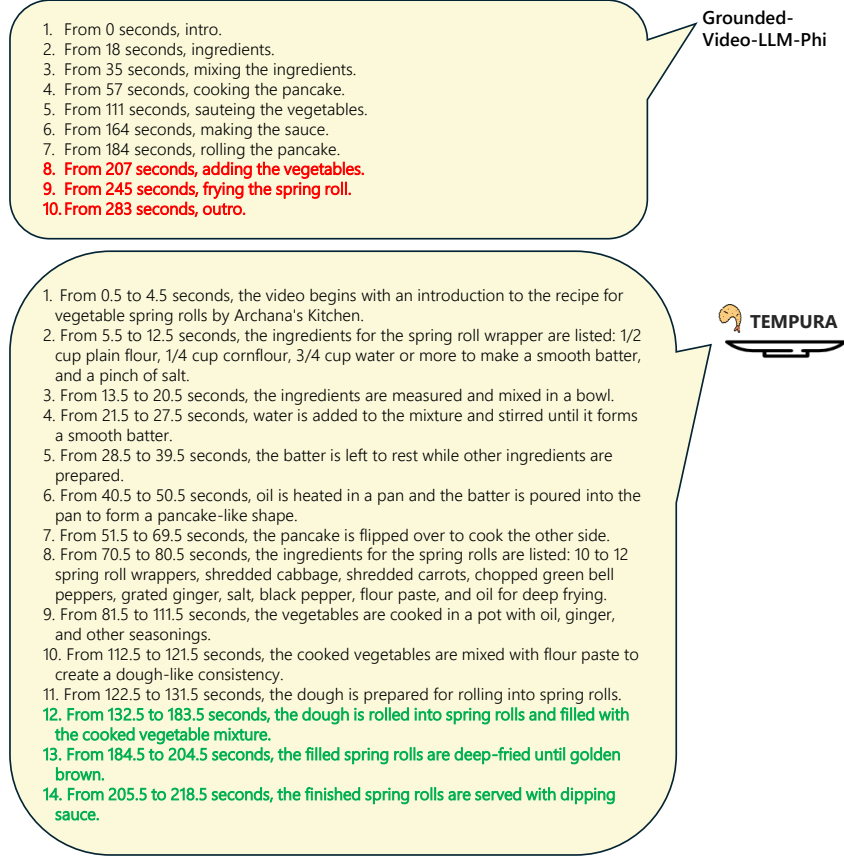


Figure 11: Comparison of **long video temporal grounding responses** on a cooking tutorial video, generated by TEMPURA (our model) and Grounded-Video-LLM-Phi (Wang et al., 2025) (baseline model). Each of the models is prompted to segment videos and describe each of the video segments in detail with correct start and end times. The text highlighted in red indicates incorrect determination of start/end times and video segment descriptions. The text in green indicates correct determination of start/end times and video segment descriptions.

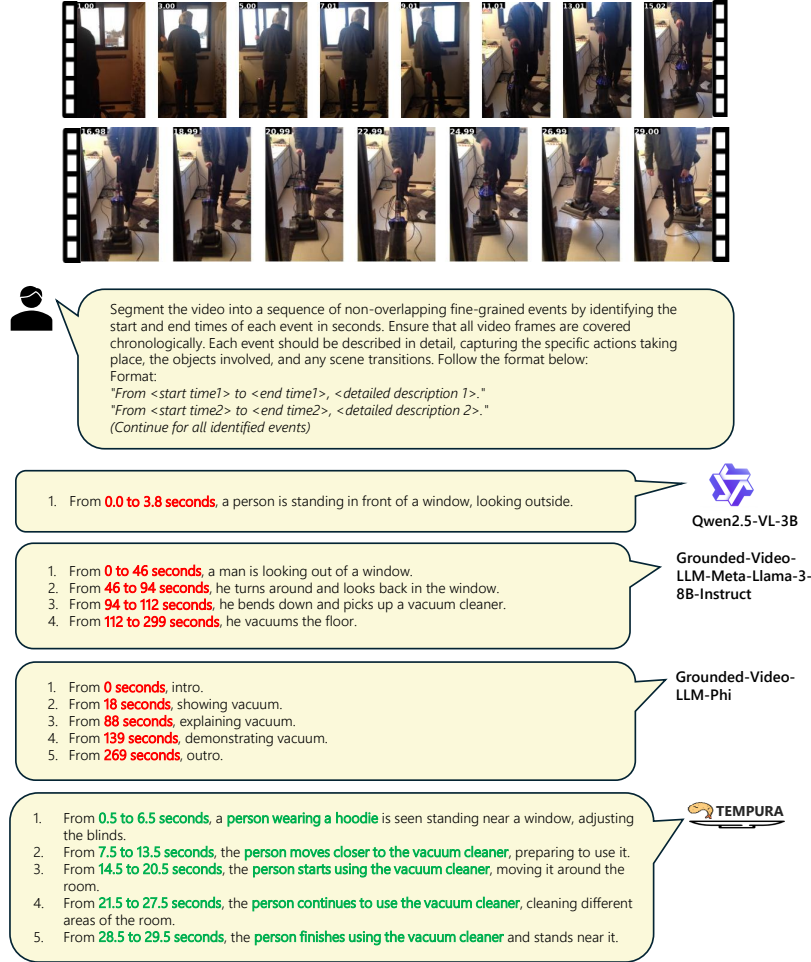


Figure 12: Comparing TEMPURA (our model) and other baseline model (Qwen2.5-VL-3B (Bai et al., 2025), Grounded-Video-LLM-Meta-Llama-3-8B-Instruct (Wang et al., 2025), and Grounded-Video-LLM-Phi (Wang et al., 2025)) abilities to generate detailed descriptions on **fine-grained events on short videos**. Each model is prompted to segment the video into fine-grained events and describe the events in detail with correct start/end timestamps. Red text indicates incorrect responses with incorrect start/end timestamps or poor descriptions of the event segment.