T2SGrid: Temporal-to-Spatial Gridification for Video Temporal Grounding

Anonymous CVPR submission

Abstract

Video Temporal Grounding (VTG) aims to localize the video segment that corresponds to a natural language query, which requires a comprehensive understanding of complex temporal dynamics. Existing Vision-LMMs typically perceive temporal dynamics via positional encoding, text-based timestamps, or visual frame numbering. However, these approaches exhibit notable limitations: assigning each frame a text-based timestamp token introduces additional computational overhead and leads to sparsity in visual attention, positional encoding struggles to capture absolute temporal information, and visual frame numbering often compromises spatial detail. To address these issues, we propose Temporal to Spatial Gridification (T2SGrid), a novel framework that reformulates video temporal understanding as a spatial understanding task. The core idea of T2SGrid is to process video content in clips rather than individual frames. we employ a overlapping sliding windows mechanism to segment the video into temporal clips. Within each window, frames are arranged chronologically in a row-major order into a composite grid image, effectively transforming temporal sequences into structured 2D layouts. The gridification not only encodes temporal information but also enhances local attention within each grid. Furthermore, T2SGrid enables the use of composite text timestamps to establish global temporal awareness. Experiments on standard VTG benchmarks demonstrate that T2SGrid achieves superior performance.

1. Introduction

Video content has become a ubiquitous medium for information dissemination, yet efficiently localizing specific moments within vast, unstructured video streams remains a fundamental challenge. Video Temporal Grounding (VTG) [4, 14, 32, 42, 43], the task of identifying the precise video segment that semantically corresponds to a natural language query, serves as a key step toward bridging this gap. Success in VTG hinges on a model's ability to comprehend not only static visual content but also intricate temporal dynamics, including action sequences, event duration, and long-range dependencies.

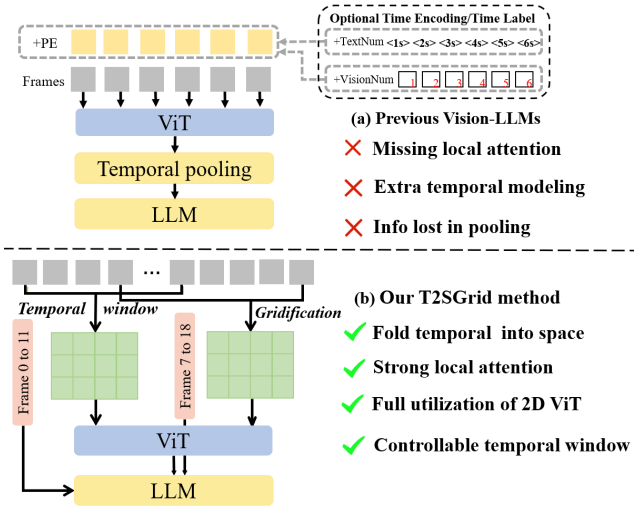

Figure 1. **Comparison between T2SGrid and previous Vision-LLMs.** (a) Traditional methods process frames sequentially or apply temporal pooling to capture information at multiple scales. However, sequential processing and pooling can cause information loss, obscure local temporal details within a window, and often require additional time encoding/label for temporal modeling. (b) Our T2SGrid method folds multiple frames within a temporal window into the spatial dimension via gridification, allowing direct processing a temporal window by the standard 2D ViT. This leverages the model's strong spatial reasoning capability for temporal understanding.

The emergence of Vision-Large Language Models (Vision-LMMs) [2, 17, 22, 51, 53] has revolutionized visual understanding [50], exhibiting remarkable zero-shot reasoning [21, 25, 27, 33, 43, 56, 63, 69, 75, 79, 80] and multi-modal comprehension [13, 30, 38, 50, 76], primarily on static images. However, extending these spatially-oriented architectures to the temporal domain remains non-trivial. As shown in Figure 1 (a), Current approaches to incorporate temporal awareness [9, 18, 24, 29] include adding positional encoding (+PE) [15, 23, 35, 48, 57], as seen in Qwen2.5-VL [2] models, and using text-based timestamps

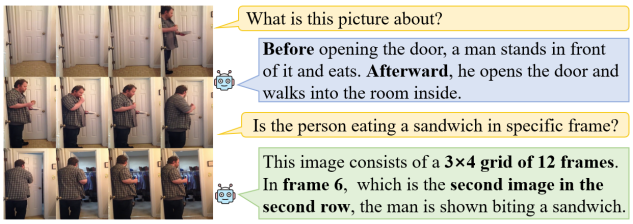

Figure 2. **Illustration of Qwen2-VL temporal reasoning on 2D grid layouts.** The model correctly infers the temporal order of events (before and after) and accurately identifies the biting action in frame 6.

(+TextNum) [31, 81, 82], as in Qwen3-VL [51] models, and visual numbering (+VisualNum) [5, 60, 61]. However, these approaches faces inherent limitations. Positional encoding, though effective for sequence modeling, fail to capture absolute temporal positions essential for grounding specific events and require additional encoding modules. Text-based timestamps (e.g., prompting with "Frame 1" or "1 seconds") inevitably introduce a rapidly growing amount of textual tokens that leads to increasingly sparse visual attention as the video length increases. Visual frame numbering, which overlays timestamps directly on frames, degrades spatial detail and undermines the very visual features that Vision-LMMs rely on for semantic understanding.

To address these challenges, we propose Temporal to Spatial Gridification (T2SGrid), a novel framework that partially reformulates temporal reasoning as a spatial problem. Previously, Vision-LLMs treat a video as a linear sequence of frames. T2SGrid takes a different approach, which partitions the video into configurable temporal windows and arranges the frames within each window into a composite image with grid-structured layout (Figure 1 (b)), which we call "gridification" of the temporal window. Then, the Vision-LLM processes the grid images rather than the original frames.

The gridification process is central to our approach, establishing a novel paradigm for video input representation. T2SGrid offers several distinct advantages. First, it transforms temporal dynamics within a window into a spatial layout, leveraging spatial attention to enhance the understanding of local temporal dynamics. Second, the row-major frame arrangement in the grid image and the partial overlap between adjacent grid images implicitly function as a form of positional encoding. Crucially, the effectiveness of these advantages is fundamentally enabled by the demonstrated capability of modern Vision-LLMs to interpret grid-based imagery. For instance, as shown in Figure 2, these models can infer relative temporal relationships (e.g., "before" and "after") by reading the spatial configuration of the grid from top-left to bottom-right, and can also identify the specific order of frames within the grid.

As shown in Figure 1 (b), beyond the input gridification, to enhance the model's awareness of global time, we interleave the grid images with text-based timestamps. Unlike existing approaches that assign a timestamp to each individual frame, we associate a single composite text-based timestamp (e.g., "Frame 0 to 11") with each grid image. By grouping multiple frames under one temporal descriptor, the model learns to associate a local window of visual content with a unified time interval, further improving temporal understanding.

In summary, our primary contributions are as follows. (1) We introduce T2SGrid, a novel paradigm that shifts video processing from individual frames to local temporal clips by transforming sequences of frames within a sliding window into a single, composite grid image. (2) Instead of assigning a timestamp to each frame, we use a single composite text timestamp for each grid image, enhancing global temporal awareness. (3) Extensive experiments on standard VTG and VQA benchmarks demonstrate that T2SGrid achieves superior performance.

2. Related Work

2.1. Video Temporal Grounding By Vision-LLMs

Video temporal grounding (VTG) [6, 8, 34, 36, 40, 66, 74, 77] aims to locate the precise time of specific actions or events within a video. For current Vision-LLMs [28, 38, 47, 51, 53], VTG is crucial for both temporal and spatial understanding. Many Vision-LLM-based approaches [12, 20, 32, 41, 43, 52, 62, 65, 80] have been proposed to tackle the VTG task. Some methods focus on fine-tuning existing models by providing textual timestamps [19, 36, 55] and constructing large temporally annotated datasets [45, 49]. Others design specialized temporal-aware modules using pooling operations [24, 38, 64, 72, 73] to enhance temporal reasoning. There are also approaches [59, 68, 70, 78, 83] that overlay frame indices [11, 37, 46, 58, 67] onto the original images to provide temporal information and feed the frames sequentially into the model. Although these methods achieve some success, they either require designing task-specific modules for temporal reasoning or compromise spatial information to preserve temporal cues. To address these limitations, we propose T2SGrid, which employs a gridification strategy along with a lightweight textual timestamps to provide global temporal context, eliminating the need for specialized module design or extensive dataset construction, and achieving superior performance.

2.2. Grid-Based Video Representation

Although several studies [3, 7, 10, 26, 71] have explored merging multiple video frames into a single composite image for video understanding, most of them primarily target

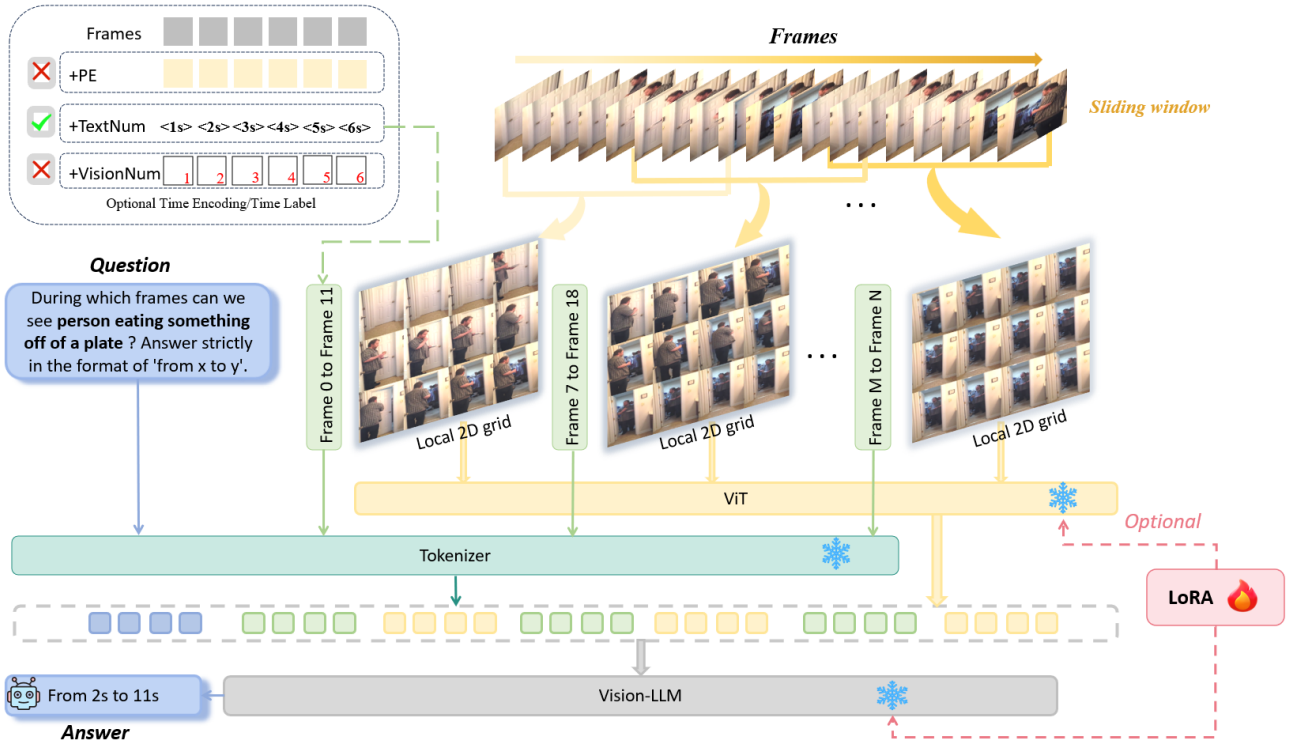

Figure 3. **Overview of our T2SGrid framework.** The original video frames are first arranged into a 2D grid in a row-major order (*gridification*) to enable spatialized temporal reasoning. A lightweight composite textual timestamp is incorporated to provide global temporal awareness. Our framework can operate in a training-free manner or be further enhanced via LoRA fine-tuning.

video question answering (VQA) [1, 39, 47]. For example, IG-VLM [26] selects a small set of keyframes and concatenates them into a single large image to facilitate VQA, yet such a keyframe grid can only model coarse-grained events. DynImg [3] enlarges a representative keyframe and appends several smaller frames below it as temporal cues to enhance video comprehension; however, such heuristic designs still fail to preserve fine-grained temporal order, thereby limiting the potential of grid-based representations for temporal reasoning. In contrast, our work is the first to reveal and leverage the intrinsic temporal information encoded within the spatial grid itself, and to apply this property to video temporal grounding, thereby enhancing the generalization capability of gridification for temporal understanding.

3. Method

We propose an innovative temporal-to-spatial gridification (T2SGrid) method. In Section 3.1, we first analyze how the gridification mechanism enhances local attention from a spatial perspective and improves temporal perception from the viewpoint of temporal attention. As illustrated in Figure 3, our approach consists of two main stages: (1) Sliding-Window Spatiotemporal Gridification, and (2) Temporal Modeling with T2SGrid, which are detailed in Section 3.2 and Section 3.3, respectively.

3.1. Attention Analysis

Current Vision-LLMs process videos by encoding frames into a sequential feature representation. This sequence, concatenated over time, is then aligned with a language query to facilitate understanding. While this approach is effective for content recognition, it presents inherent challenges for tasks requiring fine-grained temporal reasoning. To investigate the underlying limitations of this sequential methodology, we conducted an in-depth analysis of the model's internal attention mechanisms. Our central hypothesis is that the sequential processing of individual frame features biases Vision-LLMs towards recognizing a series of static spatial configurations rather than understanding the dynamic temporal evolution between them. To test this, we used Qwen2-VL-7B as case studies.

We visualize the cross-attention maps between the visual tokens and the query text token, and project them back to the original image or the grid representation. As shown in Figure 4, when given the query "a person is putting a picture onto the wall", the attention maps under sequential-frame

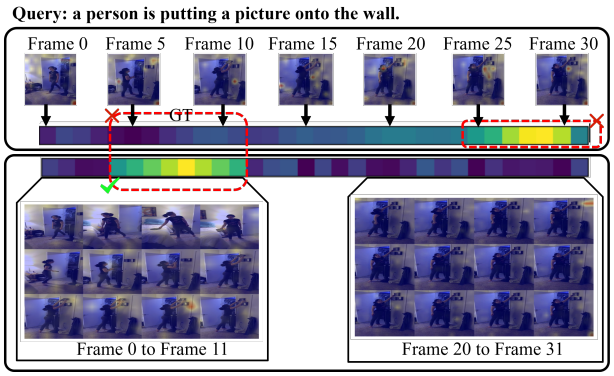

Figure 4. **Top**: spatial and temporal attention with sequential frame input. **Bottom**: spatial and temporal attention with grid-based input. The grid method captures dynamic actions and maintain focus on the correct temporal intervals.

input appear scattered or focus mainly on the person or the picture in a specific frame. This pattern indicates that sequential processing encourages the model to recognize what objects are present, but fails to capture how these objects change over time, resulting in weak sensitivity to motion-induced spatial variations. In contrast, our T2SGrid representation reorganizes the video into a structured 2D layout that places temporally adjacent patches into spatially coherent neighborhoods. This gridified structure strengthens local spatial attention and enables the model to capture subtle structural changes in the picture frame across neighboring cells. As shown in the Figure 4, the model now attends to how the person moves the picture onto the wall, demonstrating improved temporal understanding through spatial reasoning. We further average the attention over all visual tokens within each frame to obtain frame-level temporal attention. As shown in the figure, the temporal attention derived from sequential-frame input exhibits peaks that deviate substantially from the Ground Truth (GT), indicating that its temporal modeling is primarily driven by static object saliency rather than the actual progression of the action. In contrast, T2SGrid makes temporal cues more explicit: by embedding temporal adjacency into a 2D spatial layout and combining it with lightweight text-based temporal anchors, the model produces temporally consistent activation patterns that closely align with the GT. Further illustrative examples can be found in the Appendix.

In conclusion, our analyses reveal a fundamental limitation in sequential-frame processing: although the model is proficient at spatial recognition, this ability does not naturally extend to capturing fine-grained temporal progression. Sequential alignment encourages frame-by-frame object matching, causing the model to rely on static object saliency while overlooking subtle inter-frame motion cues. By contrast, the proposed gridification mechanism strengthens local spatial attention, enhances sensitivity to dynamic changes, and produces temporally consistent activation patterns that align closely with the GT. By reframing temporal reasoning into a structured 2D spatial layout, our approach enables the model to leverage its inherent spatial priors to capture genuine temporal cues and achieve more accurate action localization.

3.2. Sliding Window Spatiotemporal Gridification

We first employ a sliding window mechanism to obtain frame sequences, after which we apply spatiotemporal gridification to convert them into a coherent 2D spatial layout. Concretely, given a video with T frames, we define a temporal window size k and stride s. The i-th window W_i consists of:

$$W_i = \{f_{i\times s}, f_{i\times s+1}, \ldots, f_{i\times s+k-1}\}. \quad (1)$$

For each W_i, we spatially rearrange its k **original resolution frames** into a single composite grid G_i. Notably, our method does not compromise spatial resolution; it only performs gridification of the original frames. The layout of this grid is flexible (e.g., $M \times N$) as long as $M \times N = k$. For instance, nine frames can be arranged in a 3×3 grid. The frames are consistently ordered in a row-major fashion (left-to-right, top-to-bottom).

Temporal localization methods processing input videos that uniformly sample frame sequences are fundamentally equivalent to sliding window processing with stride $s = 1$ and window size $k = 1$. While capturing per-frame static information, this approach poorly perceives dynamic changes between adjacent frames. Our method overcomes this limitation by setting $k > 1$ (e.g., 3×3 grid), where each window contains k consecutive frames forming a local temporal segment. When processing center frame f_t, the model simultaneously receives its neighborhood $f_{t-4}, \ldots, f_{t+4}$, establishing complete dynamic context. To avoid splitting critical actions in short videos, we introduce overlap by setting the sliding-window stride $s < k$, preserving temporal continuity and allowing the model to capture both local dynamics and maintain global consistency. For longer videos, we set $s = k$ to avoid excessive overlap and computational redundancy. This design makes our approach suitable for long videos and robust to a wide range of frame rates (FPS), while maintaining the original video resolution throughout the spatiotemporal gridification process.

3.3. Temporal Modeling with T2SGrid

3.3.1. Implicit Temporal Encoding via Grid Layout

Although the T2SGrid is inherently spatial, arranging frames in a **row-major order** induces a deterministic mapping that implicitly encodes temporal information. Let N_c denote the number of frames per row. The temporal index t_f of a frame with row-column coordinates (r_f, c_f) is uniquely

defined as

$$t_f = r_f \times N_c + c_f, \tag{2}$$

reflecting the linear progression along the row-major layout.

The self-attention mechanism, however, operates on patch-level coordinates (r_p, c_p) and their corresponding 2D positional embeddings $E(r_p, c_p)$. A frame-level coordinate can be recovered from patch-level coordinates using

$$r_f = \lfloor r_p / h_{\text{patch}} \rfloor, \quad c_f = \lfloor c_p / w_{\text{patch}} \rfloor, \tag{3}$$

where h_{patch} and w_{patch} are the height and width of a frame in patches. Substituting Equation (3) into Equation (2) gives

$$t_f = \lfloor r_p / h_{\text{patch}} \rfloor \times N_c + \lfloor c_p / w_{\text{patch}} \rfloor, \tag{4}$$

showing that the temporal index is a well-defined function of patch coordinates.

Therefore, the 2D positional embeddings $E(r_p, c_p)$ contain sufficient information for the model to infer the **temporal order** of frames, enabling implicit temporal reasoning without requiring explicit timestamps or frame identifiers.

Figure 2 shows concrete examples of how Qwen2-VL [53] interpret the grid structure, providing empirical evidence that the grid layout enables implicit temporal encoding via spatial understanding.

3.3.2. Absolute Global Temporal Awareness.

Although the sliding window strategy enhances local temporal reasoning through spatiotemporal grid representations of videos, it loses the absolute temporal position of each segment. This limitation arises from its capture of only relative temporal relationships, such as action continuity between adjacent frames. This limitation prevents the establishment of absolute positioning along the global timeline, as the model loses perception of corresponding absolute time intervals in the original video (e.g., "From Frame 0 to Frame 8") while comprehending dynamic evolution within grid G_i. Such deficiency impedes video temporal localization tasks demanding precise timestamp outputs (e.g., "Xs to Ys") or global narrative comprehension (e.g., "What happened at 7 seconds in the video?"), where absolute timing remains essential to satisfy rigid requirements for a temporal reference system.

To preserve global time awareness, we introduce Absolute Time as solution. Before passing each grid image G_i to the LMM, we prepend a textual timestamp to the input prompt:

$$\text{Prompt}_i = \big[\text{“ from } T_{\text{start}} \text{ to } T_{\text{end}}\text{.”}\big]; \big[\text{Image: } G_i\big] \tag{5}$$

where T_{start} and T_{end} correspond to the start and end time of the i-th grid.

For holistic video understanding, multiple grids are organized in an interleaved text-image sequence:

$$\begin{aligned} &\big[\text{Text: } T_{\text{start}}^{(1)} \text{ to } T_{\text{end}}^{(1)}\big] \rightarrow \big[\text{Image: } G_1\big]; \\ &\big[\text{Text: } T_{\text{start}}^{(2)} \text{ to } T_{\text{end}}^{(2)}\big] \rightarrow \big[\text{Image: } G_2\big]; \\ &\qquad\vdots \\ &\big[\text{Text: } T_{\text{start}}^{(n)} \text{ to } T_{\text{end}}^{(n)}\big] \rightarrow \big[\text{Image: } G_n\big]. \end{aligned} \tag{6}$$

These absolute timestamps form a continuous temporal chain across the video, enabling the model to establish sequential relationships while preserving temporal coherence. By interleaving grids with global textual time and leveraging cross-attention mechanisms, the model can reason not only about the dynamics within each grid but also about their placement along the full video timeline.

4. Experiment

We evaluate our model on standard Video Temporal Grounding benchmarks. Following prior work [19, 42, 44, 54], we evaluate our method on the test sets of Charades-STA [14] and ActivityNet [4]. The evaluation metrics include mean Intersection over Union (mIoU) and Recall@1 at multiple IoU thresholds (R@m), where $m \in \{0.3, 0.5, 0.7\}$, consistent with previous studies [19, 44].

4.1. Experiment setup

Training dataset and benchmark We adopt a mixed training set, combining the training splits of ActivityNet-Captions [4] and Charades [14], containing 18K videos and 50K question–answer pairs. Each video in our dataset is augmented and arranged using our T2SGrid method. The question–answer pairs follow a consistent template: questions are formatted as "During which frames can we see query?", and answers are formatted as "From x to y", where x and y denote the start and end frame numbers of the queried event.

Training Detail We use Qwen2-VL-7B as the base model, which lacks temporal encoding and therefore enables a fair comparison of temporal strategies. The model is trained for 3 epochs with a batch size of 32 and a learning rate of 2×10^{-5}. We adopt LoRA fine-tuning (rank 64, $\alpha = 128$) on all linear layers. All experiments are conducted on 4×A100 GPUs. Charades-STA is trained and evaluated under the overlap configuration (g43_s7), whereas ActivityNet-Captions uses the no-overlap configuration (g43_s12). Additional details are provided in the Appendix.

4.2. Main result

4.2.1. Superior performance on VTG datsets

Table 1 reports the performance of our T2SGrid and T2SGrid-FT methods compared with previous approaches

Table 1. **Performance comparison on the video temporal grounding task with prior state-of-the-art methods.** *T2SGrid* denotes the use of spatiotemporal gridification with global time awareness, while *T2SGrid-FT* indicates fine-tuning with the instruction dataset augmented using our T2SGrid method.

Model	Charades-STA				ActivityNet			
	R@0.3	R@0.5	R@0.7	mIoU	R@0.3	R@0.5	R@0.7	mIoU
VTG-Tuned Vision-LLMs								
GroundingGPT [32]	-	29.6	11.9	-	-	-	-	-
LITA [20]	-	-	-	-	-	25.9	-	28.6
VTG-LLM [17]	52.0	33.8	15.7	-	-	-	-	-
TimeChat [44]	47.7	22.9	12.5	30.6	30.2	16.9	8.2	21.8
VTimeLLM [19]	51.0	27.5	11.4	31.2	44.0	27.8	14.3	30.4
Momentor [42]	42.9	23.0	12.4	29.3	42.6	26.6	11.6	28.5
HawkEye [54]	50.6	31.4	14.5	33.7	49.1	29.3	10.7	32.7
TimeSuite [72]	69.9	48.7	24.0	-	-	-	-	-
TRACE [16]	-	40.3	19.4	-	-	37.7	24.0	39.0
NumPro [61]	63.8	42.0	20.6	41.4	55.6	37.5	20.6	38.8
General Vision-LLMs								
GPT-4o [22]	55.0	32.0	11.5	35.4	33.3	21.2	10.4	23.7
+*T2SGrid*	57.3 (+2.3)	36.7 (+4.7)	14.8 (+3.3)	36.9 (+1.5)	47.5 (+14.2)	32.1 (+10.9)	19.7 (+9.3)	35.6 (+11.9)
Qwen3-VL-8B [51]	69.3	43.4	17.5	43.1	39.0	26.2	15.8	29.3
+*T2SGrid*	71.4 (+2.1)	47.0 (+3.6)	20.7 (+3.2)	44.9 (+1.8)	44.8 (+5.8)	26.8 (+0.6)	16.1 (+0.3)	32.5 (+3.2)
LLaVA-OneVision-1.5-8B [28]	19.8	6.7	2.3	14.5	9.2	4.9	2.1	7.5
+*T2SGrid*	45.0 (+25.2)	26.3 (+19.6)	11.9 (+9.6)	28.8 (+14.3)	31.5 (+22.3)	17.4 (+12.5)	10.4 (+8.3)	23.6 (+16.1)
Qwen2-VL-7B [53]	8.7	5.4	2.4	7.9	17.0	9.4	3.9	12.5
+*T2SGrid*	70.1 (+61.4)	46.7 (+41.3)	20.1 (+17.7)	44.3 (+36.4)	46.2 (+29.2)	27.2 (+17.8)	15.4 (+11.5)	33.3 (+20.8)
+*T2SGrid-FT*	**76.9** (+68.2)	**60.6** (+55.2)	**35.9** (+33.5)	**53.2** (+45.3)	**64.4** (+47.4)	**48.4** (+39.0)	**29.5** (+25.6)	**46.7** (+34.2)

on video temporal grounding benchmarks. After integrating the proposed T2SGrid temporal encoding, both open-source and closed-source models exhibit consistent improvements.

Specifically, GPT-4o, which already possesses strong temporal reasoning capability, achieves further gains with T2SGrid, showing that our method can enhance well-trained multimodal LLMs. Qwen3-VL-8B, a state-of-the-art vision-LLM utilizing textual timestamps for temporal modeling, also benefits modestly; for instance, its Charades-STA R@0.3 increases from 69.3 to 71.4. The smaller gain mainly arises from a slight conflict between our local temporal encoding and its original timestamp-based scheme. In contrast, Qwen2-VL-7B, which lacks explicit temporal encoding, exhibits a significant performance boost when combined with T2SGrid, reaching 70.1 R@0.3 and 44.3 mIoU on Charades-STA thatsurpassing several VTG-tuned video LLMs.

Finally, LLaVA-OneVision1.5-8B, a model trained solely on static images, achieves remarkable improvements with T2SGrid: absolute gains of 25.2, 19.6, 9.6, and 14.3 on R@0.3, R@0.5, R@0.7, and mIoU, respectively. On ActivityNet, the corresponding improvements are 22.3, 12.5, 8.3, and 16.1. These results confirm that T2SGrid effectively leverages the spatial reasoning capability of image-based models to enable temporal understanding.

To further validate the effectiveness of our approach, We fine-tune Qwen2-VL-7B on a dataset augmented and arranged using our T2SGrid framework, resulting in the best mIoU scores on Charades-STA and ActivityNet, reaching 53.2 and 46.7, respectively.

4.2.2. Training Effectiveness

In Table 5, we observe consistent performance gains when incorporating T2SGrid during fine-tuning (+T2SGrid-FT). Both the standard fine-tuning (FT) and T2SGrid-FT experiments are conducted on the same training data, ensuring a fair comparison. For LLaVA-OneVision-1.5-8B, T2SGrid yields clear improvements of +4.8 R@0.5 and +2.4 mIoU on Charades-STA, and large boosts of +13.4 R@0.3 and +10.2 mIoU on ActivityNet. For Qwen2-VL-7B, it further improves results by +3.8 R@0.5 and +2.8 mIoU on Charades-STA, and by +9.8 R@0.5, +10.3 R@0.7, and +9.0 mIoU on ActivityNet. Overall, models fine-tuned with our T2SGrid framework not only achieve consistently higher scores across all metrics and datasets but also demonstrate stronger temporal grounding capability, confirming the superior training effectiveness of our approach.

4.2.3. Qualitative Results

Figure 5 illustrates a visualization on the Charades dataset, comparing our method with previous approaches. Given the query "person throwing a blanket onto the vacuum", other methods predict incorrect temporal intervals, whereas our

Table 2. **Training effectiveness of our method.** T2SGrid-FT consistently outperforms standard fine-tuning (FT) when trained on the same data with identical training settings, demonstrating the benefits of our approach.

Model	Charades-STA				ActivityNet			
	R@0.3	R@0.5	R@0.7	mIoU	R@0.3	R@0.5	R@0.7	mIoU
LLaVA-OneVision-1.5-8B [28]	19.8	6.7	2.3	14.5	9.2	4.9	2.1	7.5
+*FT*	70.6	52.9	29.8	48.4	50.1	35.1	18.5	33.4
+*T2SGrid-FT*	73.9 (+3.3)	57.7 (+4.8)	32.7 (+2.9)	50.8 (+2.4)	63.5 (+13.4)	45.4 (+10.3)	26.2 (+7.7)	44.6 (+10.2)
Qwen2-VL-7B [53]	8.7	5.4	2.4	7.9	17.0	9.4	3.9	12.5
+*FT*	73.1	56.8	32.7	50.4	54.9	37.2	19.2	37.7
+*T2SGrid-FT*	76.8 (+3.7)	60.6 (+3.8)	35.9 (+3.2)	53.2 (+2.8)	64.4 (+9.5)	48.4 (+9.8)	29.5 (+10.3)	46.7 (+9.0)

Table 3. **Performance comparison on Video-MME, MVBench, and VideoInstruct benchmarks**. Our T2SGrid method consistently improves temporal perception and action understanding, demonstrating strong generalization QA tasks.

Model	Video MME			MVbench					VideoInstruct	
	Temporal Perception	Temporal Reasoning	All	Action Sequence	Fine-grained Action	State Change	Scene Transition	All	Temporal Understanding	Detail Orientation
Qwen2-VL-7B	60.0	41.7	63.3	60.5	46.5	44.0	79.5	51.7	2.47	2.57
+T2SGrid	**74.5**	**50.2**	**64.1**	**71.5**	**49.5**	**51.5**	**89.0**	**58.3**	**2.52**	**2.69**

method accurately captures the start and end times of the action, demonstrating the effectiveness of T2SGrid. Additional examples are provided in the Appendix.

4.2.4. Effectiveness on Long-Video and VQA Tasks

To further evaluate the effectiveness of our T2SGrid method in enhancing temporal understanding and action reasoning, we conduct experiments on three video understanding benchmarks: VideoMME [13], MVBench [30], and VideoInstruct [38]. Among them, VideoMME is a long-form video QA benchmark that further demonstrates our method's ability to reason over long-duration videos.

As shown in Table 3, the T2SGrid-enhanced model consistently outperforms the base Qwen2-VL-7B model across all metrics. On VideoMME, it improves temporal perception and temporal reasoning by 14.5 and 8.5, respectively. Similarly, on MVBench, TGrid yields consistent gains across all subcategories including action sequence, fine-grained action, state change, and scene transition, raising the overall score from 51.7 to 58.3. On the VideoInstruct dataset, our method also achieves higher scores in both temporal understanding and detail orientation, confirming that T2SGrid effectively enhances temporal reasoning and fine-grained perception across diverse video QA tasks.

These results demonstrate the generality and robustness of T2SGrid across both long and short video QA tasks, highlighting its strong ability to improve temporal perception and question answering performance.

4.3. Ablation studies

To isolate the contributions of each component of our method, we conduct a series of ablation studies on the Charades dataset using Qwen2-VL-7B.

Table 4. **Ablation study of key components:** composite textual timestamps (ComTextNum), sliding window, and grid layout. Each component contributes to improved temporal modeling and overall grounding performance.

ComTextNum	Sliding Window	Grid	R1@0.3	R1@0.5	R1@0.7	mIoU
✗	✗	✗	8.7	5.4	2.4	7.9
✓	✗	✗	53.5	23.2	7.9	32.9
✓	✓	✗	58.3	35.1	13.6	36.5
✓	✓	✓	**70.1**	**46.7**	**20.1**	**44.3**

Component Ablation. We conduct a comprehensive ablation study to evaluate the contribution of each component in our temporal encoding framework. ComTextNum denotes the use of a textual timestamp inserted before each grid element, providing an explicit temporal anchor for the model. Sliding Window refers to partitioning the sequential frames into temporally localized windows, while Grid represents the transformation of frames within each window into a 2D spatial grid.

As shown in Table 4, introducing ComTextNum substantially improves temporal information, boosting mIoU from 7.9 to 32.9. Incorporating the sliding window mechanism further enhances local spatiotemporal attention, raising mIoU to 36.5. When we additionally reformulate each temporal window into a 2D grid structure, the model achieves the largest performance gain, reaching 70.1 R1@0.3 and 44.3 mIoU. These improvements demonstrate that local implicit temporal encoding within the grid provides advantages over TextNum, validating the effectiveness of our core grid-based approach.

Temporal Modeling Method Ablation. We conduct an ablation study to evaluate four temporal encoding strategies on the Qwen2-VL-7B baseline: PE, TextNum, VisualNum,

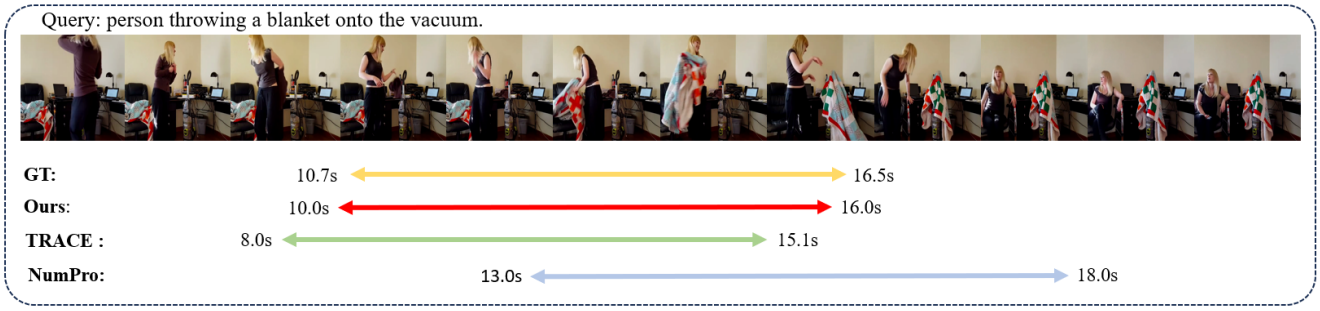

Figure 5. Qualitative Comparison with State-of-the-Art. Our method outperforms TRACE [16] and Numpro [61] on Charades by accurately identifying event boundaries in challenging scenes.

Table 5. Comparison of different temporal modeling strategies applied to the Qwen2-VL-7B baseline, including position encoding (PE), textual timestamping (TextNum), visual frame-level numeric labeling (VisualNum), and our proposed T2SGrid framework. mToken and mTime/s denote the average token count and inference time per sample, respectively.

Temporal Modeling	R1@0.3	R1@0.5	R1@0.7	mIoU	mToken	mTime/s
PE	53.1	28.4	10.6	33.8	5760.4	1.69
TextNum	53.5	23.2	7.9	32.5	5791.2	1.45
VisualNum	60.7	36.8	15.9	38.5	5760.4	2.17
Ours (wo/ overlap)	64.5	42.9	18.1	41.2	5766.1	1.43
Ours (w/ overlap)	**70.2**	**46.7**	**20.1**	**44.3**	7909.7	2.31

and Ours. PE encodes temporal order through positional encoding; TextNum adds textual timestamps before each frame or grid element; VisualNum draws numeric indices directly on frames. As shown in Table 5, Under the non-overlap setting, our method delivers the best overall performance while reducing inference time by 34.1% over VisualNum. With overlap, it trades only 6% additional time for a 14% performance gain.

Ablation Study on Grid Configuration We conduct ablations over different grid configurations, denoted as `g{col}{row}_s{stride}` where $k = \text{col} \times \text{row}$.

In Stage 1, we vary both grid size and stride (with $k = s$). As shown in Table 6, performance improves as the window size increases: R1@0.3 rises from 53.5 (g11_s1) to 63.7 (g33_s9), and mIoU increases from 32.9 to 41.2. The best configuration in this stage is g43_s12, reaching 64.5 R1@0.3 and 41.2 mIoU. However, excessively large grids (e.g., g44_s16) lead to degraded performance, with mIoU dropping to 35.9. Similarly, g43 is also the optimal configuration on ActivityNet.

Notably, under the no-overlap setting, our method does not reduce the resolution of the original frames, with the average token count (mToken) essentially the same as sequential frame input (g11_s1), since we simply concatenate full resolution frames into a single grid. This design already yields strong gains.

In Stage 2, we fix the grid size to g43 and adjust the stride

Table 6. **Ablation results of grid configurations on Qwen2-VL-7B.** `g{col}{row}_s{stride}` denoting grid size and stride. Stage 1 varies size and stride, and Stage 2 refines stride for g43.

Config	R1@0.3	R1@0.5	R1@0.7	mIoU	mToken	mTime/s
Stage 1: Exploring Grid size and stride						
g11_s1	53.5	23.2	7.9	32.9	5791.2	1.45
g22_s4	56.5	26.9	8.7	34.2	5789.0	1.43
g23_s6	61.6	38.7	16.2	38.8	5784.3	1.42
g32_s6	61.8	38.8	16.9	39.0	5784.3	1.43
g33_s9	63.7	42.8	**18.4**	41.1	5782.1	1.43
g34_s12	64.1	42.7	18.0	41.1	5781.4	1.42
g43_s12	**64.5**	**42.9**	18.1	**41.2**	5781.4	1.42
g44_s16	61.6	29.1	10.4	35.9	5779.7	1.41
Stage 2: Fine-tuning stride for optimal Grid size (g43)						
g43_s5	68.9	45.1	18.2	43.2	8971.4	2.39
g43_s6	69.5	46.4	19.6	43.7	8357.2	2.34
g43_s7	**70.2**	**46.7**	**20.1**	**44.3**	7909.6	2.31
g43_s12	64.5	42.9	18.1	41.2	5781.4	1.42

to introduce overlap. Overlapping windows yield consistent improvements: as the stride decreases from 12 to 7, R1@0.3 increases from 64.5 to 70.2, and mIoU improves from 41.2 to 44.3. The best configuration is g43_s7, which increases the number of tokens modestly but brings best performance.

5. Conlcusion

In this work, we introduced T2SGrid, a novel framework designed to tackle the challenge of video temporal grounding (VTG) for Vision-LLMs. Our core idea is to reformulate temporal understanding as a spatial reasoning problem. By "gridifying" frames from a sliding temporal window into a single composite 2D image, T2SGrid directly leverages the powerful, pre-existing spatial attention mechanisms of 2D Vision Transformers. This approach enhances the model's ability to capture fine-grained local temporal dynamics without the need for task-specific temporal modules. We further combined this local gridification with a global absolute time modeling strategy, using composite text timestamps for each grid. Together, these components provide a stronger temporal encoding mechanism. Extensive experiments on standard VTG and VQA benchmarks demonstrate that T2SGrid achieves superior performance.

References

[1] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015. 3

[2] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 1

[3] Xiaoyi Bao, Chenwei Xie, Hao Tang, Tingyu Weng, Xiaofeng Wang, Yun Zheng, and Xingang Wang. Dynimg: Key frames with visual prompts are good representation for multi-modal video understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23678–23688, 2025. 2, 3

[4] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pages 961–970, 2015. 1, 5

[5] Mu Cai, Haotian Liu, Siva Karthik Mustikovela, Gregory P Meyer, Yuning Chai, Dennis Park, and Yong Jae Lee. Vip-llava: Making large multimodal models understand arbitrary visual prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12914–12923, 2024. 2

[6] Jingyuan Chen, Xinpeng Chen, Lin Ma, Zequn Jie, and Tat-Seng Chua. Temporally grounding natural sentence in video. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 162–171, Brussels, Belgium, 2018. Association for Computational Linguistics. 2

[7] Lin Chen, Xilin Wei, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Zhenyu Tang, Li Yuan, et al. Sharegpt4video: Improving video understanding and generation with better captions. *Advances in Neural Information Processing Systems*, 37:19472–19495, 2024. 2

[8] Yi-Wen Chen, Yi-Hsuan Tsai, and Ming-Hsuan Yang. End-to-end multi-modal video temporal grounding. *Advances in Neural Information Processing Systems*, 34:28442–28453, 2021. 2

[9] Andong Deng, Zhongpai Gao, Anwesa Choudhuri, Benjamin Planche, Meng Zheng, Bin Wang, Terrence Chen, Chen Chen, and Ziyan Wu. Seq2time: Sequential knowledge transfer for video llm temporal grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13766–13775, 2025. 1

[10] Lars Doorenbos, Federico Spurio, and Juergen Gall. Video panels for long video understanding, 2025. 2

[11] Zhizhao Duan, Hao Cheng, Duo Xu, Xi Wu, Xiangxie Zhang, Xi Ye, and Zhen Xie. Cityllava: Efficient fine-tuning for vlms in city scenario, 2024. 2

[12] Zheyu Fan, Jiateng Liu, Yuji Zhang, Zihan Wang, Yi R. Fung, Manling Li, and Heng Ji. Video-llms with temporal visual screening, 2025. 2

[13] 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 Computer Vision and Pattern Recognition Conference*, pages 24108–24118, 2025. 1, 7

[14] Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. Tall: Temporal activity localization via language query. In *Proceedings of the IEEE international conference on computer vision*, pages 5267–5275, 2017. 1, 5

[15] Junqi Ge, Ziyi Chen, Jintao Lin, Jinguo Zhu, Xihui Liu, Jifeng Dai, and Xizhou Zhu. V2pe: Improving multimodal long-context capability of vision-language models with variable visual position encoding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 21070–21084, 2025. 1

[16] Yongxin Guo, Jingyu Liu, Mingda Li, Qingbin Liu, Xi Chen, and Xiaoying Tang. Trace: Temporal grounding video llm via causal event modeling. *arXiv preprint arXiv:2410.05643*, 2024. 6, 8

[17] Yongxin Guo, Jingyu Liu, Mingda Li, Dingxin Cheng, Xiaoying Tang, Dianbo Sui, Qingbin Liu, Xi Chen, and Kevin Zhao. Vtg-llm: Integrating timestamp knowledge into video llms for enhanced video temporal grounding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3302–3310, 2025. 1, 6

[18] Miran Heo, Min-Hung Chen, De-An Huang, Sifei Liu, Subhashree Radhakrishnan, Seon Joo Kim, Yu-Chiang Frank Wang, and Ryo Hachiuma. Omni-rgpt: Unifying image and video region-level understanding via token marks. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 3919–3930, 2025. 1

[19] Bin Huang, Xin Wang, Hong Chen, Zihan Song, and Wenwu Zhu. Vtimellm: Empower llm to grasp video moments, 2023. 2, 5, 6

[20] De-An Huang, Shijia Liao, Subhashree Radhakrishnan, Hongxu Yin, Pavlo Molchanov, Zhiding Yu, and Jan Kautz. Lita: Language instructed temporal-localization assistant. In *European Conference on Computer Vision*, pages 202–218. Springer, 2024. 2, 6

[21] Zeyi Huang, Yuyang Ji, Xiaofang Wang, Nikhil Mehta, Tong Xiao, Donghyun Lee, Sigmund Vanvalkenburgh, Shengxin Zha, Bolin Lai, Licheng Yu, Ning Zhang, Yong Jae Lee, and Miao Liu. Building a mind palace: Structuring environment-grounded semantic graphs for effective long video analysis with llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24169–24179, 2025. 1

[22] 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. 1, 6

[23] Habib Irani and Vangelis Metsis. Positional encoding in transformer-based time series models: a survey. *arXiv preprint arXiv:2502.12370*, 2025. 1

[24] Jinhyun Jang, Jungin Park, Jin Kim, Hyeongjun Kwon, and Kwanghoon Sohn. Knowing where to focus: Event-aware

transformer for video grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 13846–13856, 2023. 1, 2

[25] Yongrae Jo, Seongyun Lee, Aiden SJ Lee, Hyunji Lee, Hanseok Oh, and Minjoon Seo. Zero-shot dense video captioning by jointly optimizing text and moment, 2023. 1

[26] Wonkyun Kim, Changin Choi, Wonseok Lee, and Wonjong Rhee. An image grid can be worth a video: Zero-shot video question answering using a vlm. *IEEE Access*, 2024. 2, 3

[27] Jin-Seop Lee, SungJoon Lee, Jaehan Ahn, YunSeok Choi, and Jee-Hyong Lee. Tag: A simple yet effective temporal-aware approach for zero-shot video temporal grounding, 2025. 1

[28] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. 2, 6, 7

[29] Haoran Li, Yingjie Qin, Baoyuan Ou, Lai Xu, and Ruiwen Xu. Hope: Hybrid of position embedding for length generalization in vision-language models. *arXiv preprint arXiv:2505.20444*, 2025. 1

[30] Kunchang Li, Yali Wang, Yinan 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*, pages 22195–22206, 2024. 1, 7

[31] Mengtian Li, Chengshuo Zhai, Shengxiang Yao, Zhifeng Xie, Keyu Chen, and Yu-Gang Jiang. Infinite motion: Extended motion generation via long text instructions. *arXiv preprint arXiv:2407.08443*, 2024. 2

[32] 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. *arXiv preprint arXiv:2401.06071*, 2024. 1, 2, 6

[33] Ruotong Liao, Max Erler, Huiyu Wang, Guangyao Zhai, Gengyuan Zhang, Yunpu Ma, and Volker Tresp. Videoinsta: Zero-shot long video understanding via informative spatial-temporal reasoning with llms. *arXiv preprint arXiv:2409.20365*, 2024. 1

[34] Kevin Qinghong Lin, Pengchuan Zhang, Joya Chen, Shraman Pramanick, Difei Gao, Alex Jinpeng Wang, Rui Yan, and Mike Zheng Shou. Univtg: Towards unified video-language temporal grounding, 2023. 2

[35] Zikang Liu, Longteng Guo, Yepeng Tang, Tongtian Yue, Junxian Cai, Kai Ma, Qingbin Liu, Xi Chen, and Jing Liu. Vrope: Rotary position embedding for video large language models, 2025. 1

[36] Weiheng Lu, Jian Li, An Yu, Ming-Ching Chang, Shengpeng Ji, and Min Xia. Llava-mr: Large language-and-vision assistant for video moment retrieval, 2024. 2

[37] Chuofan Ma, Yi Jiang, Jiannan Wu, Zehuan Yuan, and Xiaojuan Qi. Groma: Localized visual tokenization for grounding multimodal large language models, 2024. 2

[38] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad 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)*, pages 12585–12602, 2024. 1, 2, 7

[39] Juhong Min, Shyamal Buch, Arsha Nagrani, Minsu Cho, and Cordelia Schmid. Morevqa: Exploring modular reasoning models for video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13235–13245, 2024. 3

[40] Guoshun Nan, Rui Qiao, Yao Xiao, Jun Liu, Sicong Leng, Hao Zhang, and Wei Lu. Interventional video grounding with dual contrastive learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2765–2775, 2021. 2

[41] Shraman Pramanick, Effrosyni Mavroudi, Yale Song, Rama Chellappa, Lorenzo Torresani, and Triantafyllos Afouras. Enrich and detect: Video temporal grounding with multimodal llms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 24297–24308, 2025. 2

[42] 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. *arXiv preprint arXiv:2402.11435*, 2024. 1, 5, 6

[43] 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*, pages 1847–1856, 2024. 1, 2

[44] 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*, pages 14313–14323, 2024. 5, 6

[45] Kate Sanders and Benjamin Van Durme. A survey of video datasets for grounded event understanding, 2024. 2

[46] Aleksandar Shtedritski, Christian Rupprecht, and Andrea Vedaldi. What does clip know about a red circle? visual prompt engineering for vlms, 2023. 2

[47] 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*, 2025. 2, 3

[48] Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding, 2023. 1

[49] Chaolei Tan, Zihang Lin, Junfu Pu, Zhongang Qi, Wei-Yi Pei, Zhi Qu, Yexin Wang, Ying Shan, Wei-Shi Zheng, and Jian-Fang Hu. Synopground: A large-scale dataset for multi-paragraph video grounding from tv dramas and synopses. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 8383–8392, 2024. 2

[50] Yunlong Tang, Jing Bi, Siting Xu, Luchuan Song, Susan Liang, Teng Wang, Daoan Zhang, Jie An, Jingyang Lin, Rongyi Zhu, et al. Video understanding with large language models: A survey. *IEEE Transactions on Circuits and Systems for Video Technology*, 2025. 1

[51] Qwen Team. Qwen3-vl: Sharper vision, deeper thought, broader action. *Qwen Blog. Accessed*, pages 10–04, 2025. 1, 2, 6

[52] 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, 2025. 2

[53] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 1, 2, 5, 6, 7

[54] Yueqian Wang, Xiaojun Meng, Jianxin Liang, Yuxuan Wang, Qun Liu, and Dongyan Zhao. Hawkeye: Training video-text llms for grounding text in videos, 2024. 5, 6

[55] Ye Wang, Ziheng Wang, Boshen Xu, Yang Du, Kejun Lin, Zihan Xiao, Zihao Yue, Jianzhong Ju, Liang Zhang, Dingyi Yang, et al. Time-r1: Post-training large vision language model for temporal video grounding. *arXiv preprint arXiv:2503.13377*, 2025. 2

[56] Ziyang Wang, Shoubin Yu, Elias Stengel-Eskin, Jaehong Yoon, Feng Cheng, Gedas Bertasius, and Mohit Bansal. Videotree: Adaptive tree-based video representation for llm reasoning on long videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3272–3283, 2025. 1

[57] Xilin Wei, Xiaoran Liu, Yuhang Zang, Xiaoyi Dong, Pan Zhang, Yuhang Cao, Jian Tong, Haodong Duan, Qipeng Guo, Jiaqi Wang, Xipeng Qiu, and Dahua Lin. Videorope: What makes for good video rotary position embedding?, 2025. 1

[58] Junda Wu, Zhehao Zhang, Yu Xia, Xintong Li, Zhaoyang Xia, Aaron Chang, Tong Yu, Sungchul Kim, Ryan A Rossi, Ruiyi Zhang, et al. Visual prompting in multimodal large language models: A survey. *arXiv preprint arXiv:2409.15310*, 2024. 2

[59] Mingrui Wu, Xinyue Cai, Jiayi Ji, Jiale Li, Oucheng Huang, Gen Luo, Hao Fei, Guannan Jiang, Xiaoshuai Sun, and Rongrong Ji. Controlmllm: Training-free visual prompt learning for multimodal large language models. In *Advances in Neural Information Processing Systems*, pages 45206–45234. Curran Associates, Inc., 2024. 2

[60] Tung-Yu Wu, Sheng-Yu Huang, and Yu-Chiang Frank Wang. Data-efficient 3d visual grounding via order-aware referring. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 3107–3117. IEEE, 2025. 2

[61] Yongliang Wu, Xinting Hu, Yuyang Sun, Yizhou Zhou, Wenbo Zhu, Fengyun Rao, Bernt Schiele, and Xu Yang. Number it: Temporal grounding videos like flipping manga. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 13754–13765, 2025. 2, 6, 8

[62] Huijuan Xu, Kun He, Leonid Sigal, Stan Sclaroff, and Kate Saenko. Text-to-clip video retrieval with early fusion and re-captioning. *arXiv preprint arXiv:1804.05113*, 2(6):7, 2018. 2

[63] 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, 2024. 1

[64] Weijie Xu, Jingwei Tan, Shulin Wang, and Sheng Yang. Temporal relation-aware global attention network for temporal action detection. In *International Conference on Intelligent Computing*, pages 257–269. Springer, 2024. 2

[65] Yifang Xu, Yunzhuo Sun, Zien Xie, Benxiang Zhai, and Sidan Du. Vtg-gpt: Tuning-free zero-shot video temporal grounding with gpt. *Applied Sciences*, 14(5):1894, 2024. 2

[66] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Tubedetr: Spatio-temporal video grounding with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16442–16453, 2022. 2

[67] Jianwei Yang, Hao Zhang, Feng Li, Xueyan Zou, Chunyuan Li, and Jianfeng Gao. Set-of-mark prompting unleashes extraordinary visual grounding in gpt-4v. *arXiv preprint arXiv:2310.11441*, 2023. 2

[68] Lingfeng Yang, Yueze Wang, Xiang Li, Xinlong Wang, and Jian Yang. Fine-grained visual prompting, 2023. 2

[69] Zaiquan Yang, Yuhao Liu, Gerhard Hancke, and Rynson W. H. Lau. Unleashing the potential of multimodal llms for zero-shot spatio-temporal video grounding, 2025. 1

[70] Yuan Yao, Ao Zhang, Zhengyan Zhang, Zhiyuan Liu, Tat-Seng Chua, and Maosong Sun. Cpt: Colorful prompt tuning for pre-trained vision-language models. *AI Open*, 5:30–38, 2024. 2

[71] Jinhui Ye, Zihan Wang, Haosen Sun, Keshigeyan Chandrasegaran, Zane Durante, Cristobal Eyzaguirre, Yonatan Bisk, Juan Carlos Niebles, Ehsan Adeli, Li Fei-Fei, Jiajun Wu, and Manling Li. Re-thinking temporal search for long-form video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8579–8591, 2025. 2

[72] 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, 2025. 2, 6

[73] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding, 2023. 2

[74] Hao Zhang, Aixin Sun, Wei Jing, and Joey Tianyi Zhou. Temporal sentence grounding in videos: A survey and future directions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(8):10443–10465, 2023. 2

[75] Jinglei Zhang, Yuanfan Guo, Rolandos Alexandros Potamias, Jiankang Deng, Hang Xu, and Chao Ma. Vtimecot: Thinking by drawing for video temporal grounding and reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 24203–24213, 2025. 1

[76] Tao Zhang, Xiangtai Li, Hao Fei, Haobo Yuan, Shengqiong Wu, Shunping Ji, Chen Change Loy, and Shuicheng Yan. Omg-llava: Bridging image-level, object-level, pixel-level reasoning and understanding. In *Advances in Neural Information Processing Systems*, pages 71737–71767. Curran Associates, Inc., 2024. 1

[77] Zhu Zhang, Zhou Zhao, Yang Zhao, Qi Wang, Huasheng Liu, and Lianli Gao. Where does it exist: Spatio-temporal video grounding for multi-form sentences. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[78] Heng Zhao, Zhao Yinjie, Bihan Wen, Yew-Soon Ong, and Joey Tianyi Zhou. Video-text prompting for weakly supervised spatio-temporal video grounding. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 19494–19505, 2024. 2

[79] Minghang Zheng, Xinhao Cai, Qingchao Chen, Yuxin Peng, and Yang Liu. Training-free video temporal grounding using large-scale pre-trained models. In *European Conference on Computer Vision*, pages 20–37. Springer, 2024. 1

[80] Minghang Zheng, Xinhao Cai, Qingchao Chen, Yuxin Peng, and Yang Liu. Training-free video temporal grounding using large-scale pre-trained models. In *Computer Vision – ECCV 2024*, pages 20–37, Cham, 2025. Springer Nature Switzerland. 1, 2

[81] Sheng Zhou, Junbin Xiao, Qingyun Li, Yicong Li, Xun Yang, Dan Guo, Meng Wang, Tat-Seng Chua, and Angela Yao. Egotextvqa: Towards egocentric scene-text aware video question answering. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 3363–3373, 2025. 2

[82] Yunsong Zhou, Linyan Huang, Qingwen Bu, Jia Zeng, Tianyu Li, Hang Qiu, Hongzi Zhu, Minyi Guo, Yu Qiao, and Hongyang Li. Embodied understanding of driving scenarios. In *European Conference on Computer Vision*, pages 129–148. Springer, 2024. 2

[83] Kangyu Zhu, Ziyuan Qin, Huahui Yi, Zekun Jiang, Qicheng Lao, Shaoting Zhang, and Kang Li. Guiding medical vision-language models with explicit visual prompts: Framework design and comprehensive exploration of prompt variations, 2025. 2