DART: Difficulty-Adaptive Routing for Zero-Shot Video Temporal Grounding

Zhengbo Zhang[1], Mark He Huang[1], Zhigang Tu[2†], and Ming-Hsuan Yang[3]

[1] Singapore University of Technology and Design
[2] Wuhan University
[3] University of California, Merced

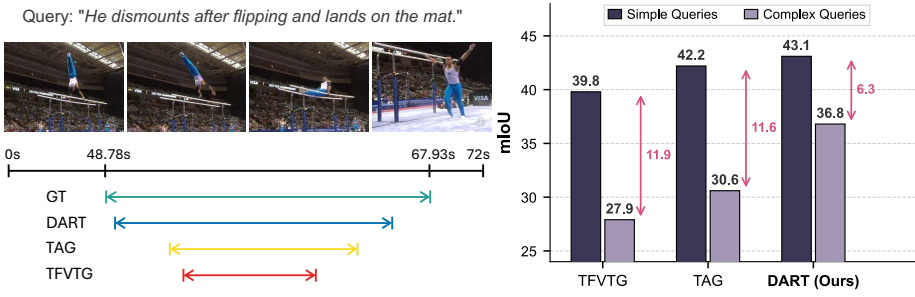

Fig. 1: Reasoning gap in zero-shot VTG. Left: a qualitative example from ActivityNet Captions [23] whose query requires temporal ordering. Feature-matching methods (TFVTG [68], TAG [26]) match only "land," missing the earlier flipping phase, while DART localizes the full event. Right: mIoU evaluated on 100 simple and 100 complex queries sampled from the ActivityNet Captions val_2 split. Feature-matching methods exhibit a large performance gap between simple and complex queries; DART reduces this gap by nearly half through difficulty-adaptive reasoning.

Abstract. Zero-shot video temporal grounding (VTG) localizes events in untrimmed videos from natural language queries without task-specific training. Existing methods rely on frame-query feature matching, which suffices for simple events but struggles with complex multi-stage queries that require understanding temporal ordering and causal structure—a disparity we call the *reasoning gap*. We propose DART (**D**ifficulty-**A**daptive **R**outing for **T**emporal Grounding), which bridges this gap by coupling difficulty-aware routing with structured reasoning in large vision-language models. A query-conditioned Determinantal Point Process (DPP) serves a dual role: selecting diverse, query-relevant keyframes as temporal evidence, and providing spectral entropy as a difficulty indicator. Simple queries are routed to a Fast path for direct prediction, while complex queries follow a Slow path with Temporal Markup Prompting, which decomposes localization into global event analysis, per-frame temporal role annotation, and boundary extraction. On Charades-STA and

† Corresponding author

ActivityNet Captions, DART achieves state-of-the-art zero-shot performance across both identically distributed and multiple out-of-distribution settings, improving mIoU by up to **3.5** points over the strongest baseline while using over 7× fewer frames. The project homepage is available at `https://dart-vtg.github.io/`.

Keywords: Video Temporal Grounding · Zero-Shot Learning · Large Vision-Language Model · Difficulty-Adaptive Reasoning

1 Introduction

Given a natural language query, localizing the target event in an untrimmed video is central to video retrieval [11], temporal question answering [28], highlight detection [27], and aligning generated motion, speech, recognizing actions under challenging illumination [9], and sound [8, 45]. This task, known as video temporal grounding (VTG) [7, 11, 23], has been extensively studied under fully supervised settings [11, 23]. However, supervised methods rely on frame-level temporal annotations, which are both expensive to collect at scale [39, 46] and inherently biased toward the training distribution [59]. Zero-shot VTG has therefore attracted growing interest. Leveraging pre-trained Vision-Language Models (VLMs) and Large Vision-Language Models (LVLMs), recent methods [20, 21, 26, 37, 51, 55, 68] score frame-query similarity via feature matching and aggregate high-scoring frames into temporal proposals, localizing events without any task-specific training data.

Although feature matching has shown promising results on simple queries, it scores each frame independently and cannot capture cross-frame dependencies such as temporal ordering, causal relationships, and event composition [29,57,68]. Complex queries whose localization demands such reasoning therefore remain a significant challenge.

To quantify this limitation, we sample 100 simple queries and 100 complex queries (the latter requiring temporal ordering, causal relations, or conditional logic) from the ActivityNet Captions [23] val_2 split and evaluate state-of-the-art zero-shot methods on each group. As shown in Fig. 1, the best-performing method drops **11.6** mIoU points on the complex group (42.2% *vs.* 30.6%). For instance, localizing *"He dismounts after flipping and lands on the mat"* requires understanding the temporal ordering of flipping, dismounting, and landing—a capability that per-frame matching lacks. We term this systematic performance gap the **reasoning gap**.

A natural way to bridge this gap is Chain-of-Thought (CoT) prompting [52], which activates the reasoning capabilities that LVLMs already possess [48] by decomposing complex predictions into explicit intermediate steps, without any task-specific training. However, applying CoT to temporal grounding is far from straightforward, introducing two key challenges. **Challenge I: How to obtain compact yet sufficient temporal evidence from raw, noisy videos for CoT reasoning?** Existing zero-shot methods typically feed uniformly sampled or fixed-stride frames to the model [37, 68], yet such query-agnostic sampling

suffers from two failure modes. First, it may miss key event stages, forcing the model to reason over incomplete temporal context [14]. Second, it retains many irrelevant frames that act as distractors, overwhelming intermediate reasoning steps and causing errors to compound along the chain [52]. **Challenge II: How to adaptively decide when reasoning is needed?** Not all queries benefit from multi-step reasoning. A simple query such as *"person sits down"* can be localized by visual matching alone; applying reasoning indiscriminately not only incurs unnecessary latency but also risks compounding errors across intermediate steps, ultimately degrading localization accuracy [35, 52].

We propose **DART** (**D**ifficulty-**A**daptive **R**outing for **T**emporal Grounding), which addresses both challenges within a single framework. The key observation is that selecting diverse, query-relevant keyframes as temporal evidence naturally exposes how many distinct stages the event contains, which in turn indicates query difficulty—thereby unifying evidence extraction and difficulty routing.

For Challenge I, we cast temporal evidence extraction as a keyframe selection problem: the selected frames should be both relevant to the query and mutually diverse, covering all stages of the target event while reducing redundancy. Jointly optimizing for relevance and diversity is non-trivial, because diversity couples every frame's marginal value to the frames already chosen, requiring in principle an exponential search over all subsets. Inspired by determinantal point processes (DPP) [24], which provide efficient greedy approximations with provable near-optimality guarantees for diverse subset selection, we design a query-conditioned variant that selects keyframes jointly maximizing query relevance and inter-frame diversity.

Our query-conditioned DPP also addresses Challenge II. Complex events that require reasoning comprise multiple sub-events with rich relational structure (*e.g.* temporal ordering, causal dependencies). Since the DPP kernel encodes this relational structure among all frames, its spectral entropy measures the effective number of dominant components, which in our VTG setting correlates with the complexity of the query-relevant temporal structure. We use spectral entropy as a difficulty indicator to route each query: queries with low spectral entropy are sent to a **Fast path** that directly predicts timestamps, while queries with high spectral entropy follow a **Slow path** that invokes our Temporal Markup Prompting, a structured CoT procedure decomposing grounding into event analysis, per-frame role tagging, and boundary extraction.

To validate the generalization and robustness of DART, we evaluate on Charades-STA [11] and ActivityNet Captions [23] under the standard IID setting, multiple OOD settings that vary temporal location bias, dataset split, and query vocabulary, and a cross-dataset transfer setting. DART sets a new state of the art across all settings, with gains of up to **3.4** mIoU on OOD and compositional splits, confirming that structured reasoning generalizes better than feature matching to novel distributions and queries.

2 Related Work

2.1 Training-based Video Temporal Grounding

Training-based Video Temporal Grounding methods can be broadly classified into three categories: fully supervised [7,11,27,62], weakly supervised [13,39], and unsupervised approaches [41]. Fully supervised VTG methods train models using pairs of natural language queries and their corresponding video segments with precisely annotated start and end boundaries [60, 64]. While such fine-grained supervision enables accurate localization, the resulting models often inherit biases from the limited and domain-specific nature of manually curated datasets, which restricts their generalization to unseen videos or domains [1, 59]. To mitigate this dependency, weakly supervised and unsupervised paradigms have been introduced [13,39,41]. Weakly supervised methods utilize coarse video-level textual descriptions without requiring frame-level temporal annotations [10, 18], whereas unsupervised approaches eliminate textual supervision entirely by generating pseudo-text queries from the visual content itself and treating them as weak labels [22, 69]. Although these approaches alleviate reliance on exhaustive human annotations, their performance remains constrained by biases in the underlying training data and the limited diversity of pseudo-labels [12, 59].

In contrast, we focus on the zero-shot setting, which aims to achieve strong generalization by harnessing the multimodal understanding and reasoning abilities of pre-trained Large Vision-Language Models (LVLMs) [30, 42, 54], without any task-specific fine-tuning [55, 56, 65–67].

2.2 Zero-shot Video Temporal Grounding

Zero-shot VTG methods [26,37,55,56,68] leverage pre-trained VLMs [30,42] and LVLMs [49,61] as feature extractors to obtain embeddings for both queries and video frames. In these frameworks, the target event is localized by computing similarity scores between textual and visual embeddings and merging temporally adjacent high-scoring frames into temporal proposals [37, 68]. Luo *et al.* [37], who first introduced the zero-shot VTG task, identify target events via similarity matching in a shared feature space, directly inferring temporal boundaries from matched frames. Recognizing that queries often describe composite events involving multiple sub-events, Zheng *et al.* [68] decompose queries into sub-event sequences, match each to corresponding visual segments, and aggregate the results to derive the final temporal span. More recently, Jeon *et al.* [21] address the granularity mismatch between queries and video content by rewriting queries at multiple granularity levels and aligning each rewritten query with corresponding video captions via CLIP-based cosine similarity. Mechanism-level explanations clarify how networks transform and route information for interpretable multimodal reasoning [2–5]. This line of work provides a valuable foundation, which we complement by using the spectrum of query-conditioned frame relations to decide when temporal reasoning is needed.

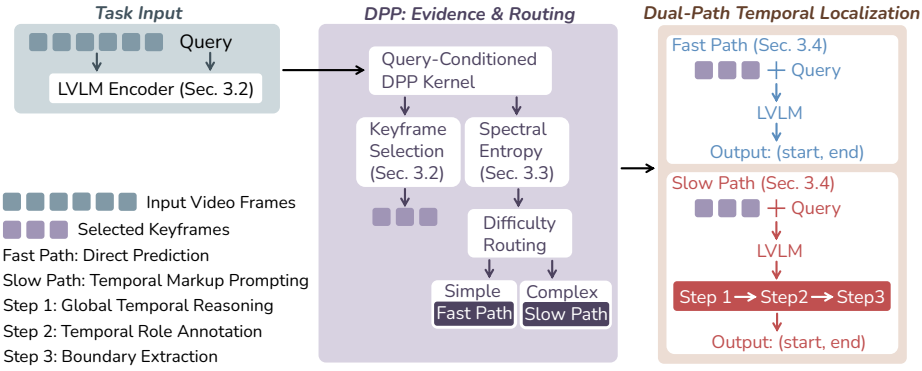

Fig. 2: Overview of the DART pipeline. The LVLM encoder denotes the vision encoder and text encoder used to extract frame features and query features, respectively. DART then (1) selects diverse, query-relevant keyframes via a DPP kernel, (2) routes each query to a fast or slow path based on spectral entropy, and (3) performs temporal localization through either direct prediction or structured reasoning.

Despite these advances, all above methods apply a fixed processing pipeline to every query regardless of its complexity. In contrast, our method introduces difficulty-adaptive routing that automatically distinguishes simple from complex queries and activates structured temporal reasoning only when needed.

3 Method

Given an untrimmed video and a natural language query, video temporal grounding (VTG) aims to predict the start and end timestamps of the segment most semantically relevant to the query.

To bridge the reasoning gap of existing feature-matching methods, we propose DART (Fig. 2), which activates the reasoning capabilities of LVLMs through three tightly coupled stages. We first construct a query-conditioned DPP kernel that jointly encodes frame-query relevance and inter-frame visual-temporal diversity, then greedily select keyframes with an adaptive stopping criterion that automatically determines the frame budget for each query (Sec. 3.2). The spectral entropy of this kernel, which reflects the relational structure among event stages, then serves as a difficulty indicator that routes each query to either direct localization or structured reasoning (Sec. 3.3). Based on this decision, the selected keyframes are passed to either direct localization or structured temporal reasoning via our Temporal Markup Prompting (Sec. 3.4). We introduce necessary background on DPP in Sec. 3.1.

3.1 Preliminaries: Determinantal Point Processes

Selecting a subset that balances element importance and inter-element diversity is computationally intractable [25]: the diversity criterion couples every element

to every other, so each element's value depends on the composition of the entire selected set, requiring in principle an exhaustive search over all possible subsets. A Determinantal Point Process (DPP) [24] provides an efficient solution: it encodes both importance and diversity into a single kernel matrix, from which a near-optimal subset can be obtained via greedy selection with provable approximation guarantees.

Specifically, given a ground set of M elements, a DPP captures element relationships through a kernel matrix $\boldsymbol{L} \in \mathbb{R}^{M \times M}$, whose diagonal entries L_{ii} encode element importance and off-diagonal entries L_{ij} encode pairwise similarity. For any subset S, the determinant $\det(\boldsymbol{L}_S)$ of the corresponding submatrix scores its quality: it measures the "volume" spanned by the selected elements, which is large when each element is important (large diagonal) and the elements are diverse (small off-diagonal). Exact maximization of $\det(\boldsymbol{L}_S)$ over all subsets is intractable, but a greedy algorithm that iteratively selects the element with the largest *marginal gain* $\Delta(e \mid S)$ achieves a near-optimal solution with provable guarantees [24]. The marginal gain measures how much adding element e to the current set S increases the log-determinant:

$$\Delta(e \mid S) = \log\det(\boldsymbol{L}_{S\cup\{e\}}) - \log\det(\boldsymbol{L}_S). \tag{1}$$

The process repeats until the desired number of elements is reached.

3.2 Preparing Temporal Evidence via Query-Conditioned DPP

Effective CoT reasoning requires selecting keyframes that are both relevant to the query and mutually diverse, covering the target event while minimizing redundancy. Since jointly optimizing both criteria requires evaluating all possible frame subsets, which is exponential in the number of frames, we leverage the DPP (Sec. 3.1) and design a *query-conditioned* variant that greedily selects diverse yet query-relevant keyframes. Specifically, it consists of two steps: (1) constructing a query-conditioned kernel that jointly encodes frame-query relevance and inter-frame visual-temporal similarity, and (2) greedily selecting keyframes from this kernel with an adaptive stopping criterion.

Query-Conditioned Kernel Construction. To jointly capture frame-query relevance and inter-frame similarity within a single optimization objective, we construct a query-conditioned kernel matrix $\tilde{\boldsymbol{L}}$. We first build a base similarity kernel $\boldsymbol{L}$ that models visual-temporal relationships among frames, then weight it by query-dependent relevance scores to obtain $\tilde{\boldsymbol{L}}$.

Specifically, given frame features $\boldsymbol{F} = [\boldsymbol{f}_1, \ldots, \boldsymbol{f}_M]$ from the vision encoder of a pretrained LVLM and a text feature $\boldsymbol{q}$ from the same LVLM's text encoder, the base similarity kernel $L_{ij} = k_v(\boldsymbol{f}_i, \boldsymbol{f}_j) \cdot k_t(t_i, t_j)$ combines a visual Gaussian kernel k_v and a temporal Gaussian kernel k_t. The product form preserves visually similar frames that are far apart in time, allowing different stages of the same event to be retained. To account for query relevance, we further weight each frame by its relevance score $w_i = \sigma(\boldsymbol{f}_i^\top \boldsymbol{q}/\tau)$, where σ is the sigmoid function and τ is a temperature parameter, and construct the query-conditioned kernel $\tilde{L}_{ij} =$

$\sqrt{w_i}\, L_{ij}\, \sqrt{w_j}$. The sigmoid maps frame-query similarity into $[0, 1]$, ensuring that query-irrelevant frames receive near-zero weight and are effectively suppressed in the kernel. The resulting $\tilde{\boldsymbol{L}}$ jointly encodes frame-query relevance (via w_i) and inter-frame diversity (via L_{ij}), enabling our query-conditioned DPP to select frames that satisfy both criteria simultaneously.

Greedy Keyframe Selection with Adaptive Budget. After obtaining $\tilde{\boldsymbol{L}}$, we greedily select keyframes by iteratively choosing the frame with the largest marginal gain $\Delta(e \mid S)$ (Eq. (1)). Since $\tilde{L}_{ij} = \sqrt{w_i}\, L_{ij}\, \sqrt{w_j}$, by the multiplicative property of determinants [24] we have $\det(\tilde{\boldsymbol{L}}_S) = (\prod_{i \in S} w_i)\, \det(\boldsymbol{L}_S)$. Substituting into Eq. (1) yields:

$$\begin{aligned}
\Delta(e \mid S) &= \log\det(\tilde{\boldsymbol{L}}_{S\cup\{e\}}) - \log\det(\tilde{\boldsymbol{L}}_S) \\
&= \log\big[\textstyle\prod_{i\in S\cup\{e\}} w_i \cdot \det(\boldsymbol{L}_{S\cup\{e\}})\big] - \log\big[\textstyle\prod_{i\in S} w_i \cdot \det(\boldsymbol{L}_S)\big] \\
&= \underbrace{\log\det(\boldsymbol{L}_{S\cup\{e\}}) - \log\det(\boldsymbol{L}_S)}_{\text{diversity}} + \underbrace{\log w_e}_{\text{query relevance}}, \qquad (2)
\end{aligned}$$

Intuitively, Eq. (2) shows that each selected frame balances two complementary objectives: the diversity term favors frames dissimilar to those already selected, while the relevance term favors frames closely related to the query.

As different events may span different numbers of stages and durations, the number of keyframes needs to vary across queries. In our method, we let the greedy process itself decide when to stop. The key observation is that Δ_k typically decreases as the greedy process proceeds: early frames capture the most informative content with large gains, while later frames contribute progressively less. We therefore stop when newly added frames contribute negligible information, *i.e.*, when the relative gain drops below a threshold: $\Delta_k/\Delta_1 < \varepsilon$, where Δ_1 is the first-step gain and ε is a hyperparameter. This way, the number of selected keyframes automatically adapts to each (video, query) pair. The greedy process outputs K keyframes paired with their timestamps as input for the LVLM.

3.3 Spectral Analysis for Difficulty Routing

After obtaining the adaptive keyframe budget for each event (Sec. 3.2), we next route each query to either direct localization or structured reasoning. A straightforward approach is to use the frame count itself as a proxy for difficulty, since events requiring more keyframes seem to be harder. However, this heuristic is unreliable: "person walks from the living room to the kitchen" may yield many keyframes because the scene changes significantly across rooms, yet the action is simple and needs no reasoning. Conversely, "person picks up a phone and answers it" may need only a few keyframes, yet involves a causal chain of distinct sub-actions that demands structured reasoning. What truly distinguishes complex queries is not the number of frames but the *relational structure* among event stages, including temporal ordering, causal transitions, and conditional dependencies (Sec. 1).

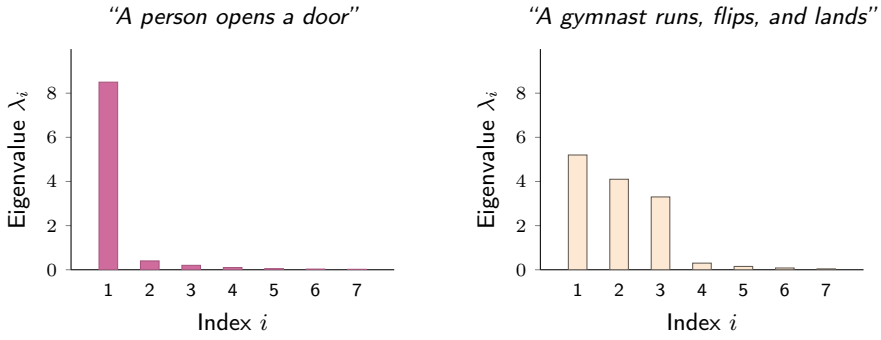

Fig. 3: Spectral contrast between simple and complex queries. Eigenvalue spectra of the query-conditioned kernel $\tilde{\boldsymbol{L}}$ for a Charades-STA query (left) and an ActivityNet Captions query (right). A simple query (left) concentrates energy in λ_1 (one event stage), yielding low $H_{\text{spectral}} \rightarrow$ **Fast path**. A complex query (right) spreads energy across λ_1–λ_3 (multiple event stages), yielding high $H_{\text{spectral}} \rightarrow$ **Slow path**.

Since our query-conditioned kernel $\tilde{\boldsymbol{L}}$ already encodes such relational structure among all frames, a natural question arises: can we read off the reasoning complexity directly from $\tilde{\boldsymbol{L}}$? Indeed, the eigenspectrum of a positive semidefinite kernel reveals how many independent components the data contains [24]: a few dominant eigenvalues indicate that the data has few independent modes, while many significant eigenvalues signal a richer multi-modal structure. Because $\tilde{\boldsymbol{L}}$ is conditioned on the query, frames unrelated to the query receive near-zero weight and are effectively suppressed in the kernel, so the eigenspectrum reflects the complexity of the query-relevant event structure rather than the general visual diversity of the video. Each dominant eigenvalue therefore corresponds to a group of query-relevant yet mutually dissimilar frames—intuitively, a distinct event stage; a simple event yields one or two dominant eigenvalues, while a complex multi-stage event spreads energy across many (Fig. 3). Building on this, we quantify the number of dominant eigenvalues via the spectral entropy of $\tilde{\boldsymbol{L}}$ to determine whether structured reasoning is needed, from a single spectral decomposition at negligible cost.

Spectral Entropy as Difficulty Indicator. As discussed above, the number of dominant eigenvalues serves as a reliable indicator that distinguishes simple from complex events. However, directly counting dominant eigenvalues requires choosing an energy threshold, and different thresholds can lead to inconsistent counts for borderline cases, making the indicator unreliable. To avoid this, we normalize the eigenvalue spectrum into a probability distribution and compute its entropy, whose information-theoretic meaning is the *effective number* of dominant eigenvalues [43], providing a smooth, information-theoretic difficulty measure that captures the complexity of the query-relevant temporal structure.

Specifically, we perform eigendecomposition of $\tilde{\boldsymbol{L}}$ to obtain its eigenvalues $\{\lambda_i\}_{i=1}^{M}$, normalize them into $q_i = \lambda_i / \sum_j \lambda_j$, and compute the spectral entropy $H_{\text{spectral}} = -\frac{1}{\log M} \sum_{i=1}^{M} q_i \log q_i \in [0, 1]$. H_{spectral} close to 0 indicates that

only one or two eigenvalues dominate (few event stages, simple query); H_{spectral} close to 1 indicates that many eigenvalues carry comparable energy (many event stages, complex query).

Difficulty-Aware Routing. Based on H_{spectral}, we route each query to one of two reasoning paths: if $H_{\text{spectral}} \leq \theta$, the query is assigned to the **Fast** path, which directly prompts the LVLM to predict start and end timestamps from the selected keyframes, sufficing for simple events; otherwise ($H_{\text{spectral}} > \theta$), it is routed to the **Slow** path, which activates our Temporal Markup Prompting (Sec. 3.4), a structured multi-step reasoning process designed for complex multi-stage events. Here θ is a hyperparameter.

3.4 Dual-Path Temporal Localization

At this point, we have determined both the keyframe set and the reasoning path for each query. For simple queries ($H_{\text{spectral}} \leq \theta$), the event structure is already clear from the selected keyframes, so the LVLM only needs to identify the matching segment. For complex queries ($H_{\text{spectral}} > \theta$), however, the event spans multiple stages with temporal and causal dependencies that cannot be resolved by visual matching alone, requiring the LVLM to reason about how the stages relate before localizing the boundaries. We design two corresponding paths to match these distinct requirements.

Fast Path: Direct Prediction. For simple queries routed to the Fast path, we provide the K selected keyframes with their timestamps to the LVLM along with a task-specific instruction that describes the temporal grounding objective and specifies the output format. The LVLM directly outputs the start and end timestamps (t_s, t_e) without any intermediate reasoning steps.

Slow Path: Temporal Markup Prompting. For complex queries, directly prompting the LVLM to output timestamps can be suboptimal, as the model needs to implicitly reason about multiple actions, their temporal ordering, and their correspondence to specific frames in a single step. Moreover, generic Chain-of-Thought prompting (*e.g.* "let's think step by step") provides no task-specific guidance and risks producing verbose, unfocused reasoning. Inspired by recent work showing that structured, task-specific prompting significantly outperforms open-ended reasoning for spatial and temporal tasks [6, 15], we design *Temporal Markup Prompting* (TMP), a structured reasoning procedure specifically tailored for temporal grounding. Unlike generic CoT, TMP prescribes a fixed three-step procedure with well-defined intermediate outputs at each step, ensuring that every step is directed toward temporal localization. The three steps are executed sequentially within a single generation:

Step 1: Global Temporal Reasoning. Before looking at any individual frame, the model is prompted to break the query into its key actions, identify how they relate in time (sequential, causal, conditional, *etc.*), and outline where the event roughly falls on the timeline. This step ensures that the model forms a global picture of the event first, so that subsequent per-frame judgments are anchored in temporal context rather than made in isolation.

Step 2: Per-Frame Temporal Role Annotation. Guided by the global analysis from Step 1, we prompt the model to assign every keyframe a temporal role from the ordered set {BEFORE, START, DURING, END, AFTER}. Two constraints enforce structured output: (i) every keyframe must receive a label, preventing the model from skipping frames or relying on implicit assumptions; (ii) labels must be monotonically non-decreasing along the timeline (BEFORE → START → DURING → END → AFTER), enforcing temporal consistency. Note that the diversity term in Sec. 3.2 ensures keyframes extend beyond event boundaries, providing context for reliable annotation.

Step 3: Boundary Extraction. The start timestamp t_s is taken from the first frame labeled START, and t_e from the last frame labeled END.

4 Experiments

4.1 Experimental Setups

Datasets. For fair comparison, we follow [68] to conduct experiments on two benchmark datasets: ActivityNet Captions [23] and Charades-STA [11]. ActivityNet Captions contains 20K videos originally collected for video captioning, with 37,417/17,505/17,031 video-query pairs in the train/val_1/val_2 splits. Following previous works [37,68], we report results on the val_2 split. Charades-STA is constructed based on the Charades dataset [44] and includes 12,408/3,720 video-query pairs in the train/test splits. We report performance on the test split.
Evaluation Metrics. We follow the evaluation metrics "R@m" and "mIoU" used in previous works [37, 68]. The metric "R@x" represents the percentage of predictions whose Intersection over Union (IoU) exceeds "x", while "mIoU" measures the average IoU across all predictions.
Implementation Details. We adopt LLaVA-1.6-7B [34] as our vision-language backbone. Each input video is downsampled to 3 FPS following TFVTG [68] to ensure consistent evaluation. For the query-conditioned DPP kernel (Sec. 3.2), we set Gaussian kernel bandwidths σ_v=0.5 for visual similarity and σ_t=2.0 for temporal distance, with temperature τ=0.1 in the relevance weight; all three are selected via grid search on the Charades-STA validation split. The adaptive stopping threshold is set to ε=0.05, and the routing threshold to θ=0.45. During inference, the LVLM uses greedy decoding with a maximum of 512 tokens. All experiments are conducted on a single NVIDIA A100 80 GB GPU.

4.2 Main Results

We follow TFVTG [68] and conduct experiments on the Charades-STA [11] and ActivityNet Captions [23] datasets under both the identically distributed (IID) setting and the out-of-distribution (OOD) setting.
IID Setting. As shown in Tab. 1, DART achieves state-of-the-art zero-shot performance on both datasets, surpassing the previous best zero-shot method TAG [26] by +3.24 mIoU on Charades-STA and +3.34 mIoU on ActivityNet

Table 1: Results under the IID setting on Charades-STA [11] and ActivityNet Captions [23]. Best zero-shot results are in **bold**, second best are underlined.

Method	Setting	Charades-STA [11]				ActivityNet Captions [23]			
		R@0.3	R@0.5	R@0.7	mIoU	R@0.3	R@0.5	R@0.7	mIoU
GroundingGPT [32]	Fully	-	29.6	11.9	-	-	-	-	-
VTimeLLM-13B [15]		55.3	34.3	14.7	34.6	44.8	29.5	14.2	31.4
2D-TAN [64]		-	39.81	23.25	-	58.75	44.05	27.38	-
EMB [16]		72.50	58.33	39.25	53.09	64.13	44.81	26.07	45.59
MGSL-Net [33]		-	63.98	41.03	-	-	51.87	31.42	-
EaTR [19]		-	68.47	44.92	-	-	58.18	37.64	-
CRM [17]	Weakly	53.66	34.76	16.37	-	55.26	32.19	-	-
CNM [70]		60.39	35.43	15.45	-	55.68	33.33	-	-
CPL [71]		66.40	49.24	22.39	-	55.73	31.37	-	-
Huang *et al.* [18]		69.16	52.18	23.94	45.20	58.07	36.91	-	41.02
Gao *et al.* [12]	Unsup.	46.69	20.14	8.27	-	46.15	26.38	11.64	-
PSVL [41]		46.47	31.29	14.17	31.24	44.74	30.08	14.74	29.62
PZVMR [47]		46.83	33.21	18.51	32.62	45.73	31.26	17.84	30.35
Kim *et al.* [22]		52.95	37.24	19.33	36.05	47.61	32.59	15.42	31.85
SPL [69]		60.73	40.70	19.62	40.47	50.24	27.24	15.03	35.44
VideoChat-7B [31]	Zero-shot	9.0	3.3	1.3	6.5	8.8	3.7	1.5	7.2
VideoLLaMA-7B [61]		10.4	3.8	0.9	7.1	6.9	2.1	0.8	6.5
VideoChatGPT-7B [38]		20.0	7.7	1.7	13.7	26.4	13.6	6.1	18.9
Luo *et al.* [37]		56.77	42.93	20.13	37.92	48.28	27.90	11.57	32.37
GranAlign [21]		59.1	39.6	22.7	38.0	50.3	**34.0**	16.5	33.1
VTG-GPT [55]		59.48	43.68	25.94	39.81	47.13	28.25	12.84	30.49
VTIMECOT [63]		66.96	38.79	20.83	43.41	-	-	-	-
TFVTG [68]		67.04	49.97	24.32	44.51	49.34	27.02	13.39	34.10
TAG [26]		67.82	48.58	26.67	45.69	51.88	28.91	15.07	36.55
Ours		**70.98**	**52.04**	**29.45**	**48.93**	**54.90**	32.14	**18.11**	**39.89**

Captions. Notably, DART also outperforms several weakly supervised methods (*e.g.*, CPL [71]) and narrows the gap with fully supervised approaches. We attribute this to the combination of query-conditioned DPP keyframe selection, which provides more informative visual input, and difficulty-adaptive routing, which applies structured CoT reasoning only when needed.

OOD Setting. Following TFVTG [68], we evaluate under moment-shifted splits where a random video of p seconds is prepended to shift ground-truth locations (OOD-1/OOD-2, Tab. 2), the distribution-shifted Charades-CD [59], and the novel-text Charades-CG [29] (Tab. 3). We also report the cross-dataset setting (Tab. 4), where supervised baselines are trained on ActivityNet Captions and tested on Charades-STA, while zero-shot methods are directly evaluated. DART achieves state-of-the-art zero-shot performance across nearly all OOD metrics, demonstrating that difficulty-adaptive reasoning generalizes better than prior feature-matching methods [37] that primarily treat LVLMs as feature extractors.

4.3 Ablation Studies

Component ablation. To verify the contribution of each component, we remove them individually and report results on Charades-STA (IID) in Tab. 6. First, replacing the query-conditioned DPP with uniform sampling (w/o DPP) causes the largest single-component drop (-4.23 mIoU), even though spectral

Table 2: Results under the OOD setting on the Charades-STA [11] and ActivityNet Captions [23] datasets.

	Method	Charades-STA [11]						ActivityNet Captions [23]					
		OOD-1			OOD-2			OOD-1			OOD-2		
		R@0.5	R@0.7	mIoU	R@0.5	R@0.7	mIoU	R@0.5	R@0.7	mIoU	R@0.5	R@0.7	mIoU
Fully	LGI [40]	42.1	18.6	41.2	35.8	13.5	37.1	16.3	6.2	22.2	11.0	3.9	17.3
	2D-TAN [64]	27.1	13.1	25.7	21.1	8.8	22.5	16.4	6.6	23.2	11.5	3.9	19.4
	MMN [50]	31.6	13.4	33.4	27.0	9.3	30.3	20.3	7.1	26.2	14.1	5.2	20.6
	VDI [36]	25.9	11.9	26.7	20.8	8.7	22.0	20.9	7.1	27.6	14.3	5.2	23.7
	DCM [58]	44.4	19.7	42.3	38.5	15.4	39.0	18.2	7.9	24.4	12.9	4.8	20.7
W.	CNM [70]	9.9	1.7	21.6	6.1	0.5	16.6	6.1	0.4	21.0	2.5	0.1	16.8
	CPL [71]	29.9	8.5	32.2	24.9	6.3	30.5	4.7	0.4	21.1	2.1	0.2	17.7
U.	PSVL [41]	3.0	0.7	8.2	2.2	0.4	6.8	-	-	-	-	-	-
	PZVMR [47]	-	8.6	25.1	-	6.5	28.5	-	4.4	28.3	-	2.6	19.1
ZS	Luo *et al.* [37]	40.3	18.2	38.2	38.9	17.0	37.8	18.4	6.8	21.1	18.6	7.4	20.6
	TFVTG [68]	45.9	20.8	43.0	43.8	20.0	42.6	20.4	11.2	31.7	18.5	10.0	30.3
	TAG [26]	45.3	23.2	44.7	44.1	22.0	44.6	28.5	14.7	36.2	28.3	14.5	36.1
	Ours	**48.7**	**25.8**	**48.1**	**47.2**	**24.5**	**47.8**	**31.8**	**17.9**	**39.7**	**31.6**	**17.6**	**39.5**

Table 3: Results on Charades-CD [59] (OOD split) and Charades-CG [29] (novel text).

	Method	Charades-CD [59]			Charades-CG [29]					
		Test-OOD			Novel-composition			Novel-word		
		R@0.3	R@0.5	R@0.7	R@0.5	R@0.7	mIoU	R@0.5	R@0.7	mIoU
Fully	2D-TAN [64]	43.45	30.77	11.75	30.91	12.23	29.75	29.36	13.21	28.47
	TSP-PRL [53]	31.93	19.37	6.20	16.30	2.04	13.52	14.83	2.61	14.03
	SCDM [60]	52.38	41.60	22.22	27.73	12.25	30.84	-	-	-
	VISA [29]	-	-	-	45.41	22.71	42.03	42.35	20.88	40.18
	DeCo [57]	-	-	-	47.39	21.06	40.70	-	-	-
W.	WSSL [10]	35.86	23.67	8.27	3.61	1.21	8.26	2.79	0.73	7.92
	CPL [71]	-	-	-	39.11	15.60	35.53	45.90	22.88	-
U.	SPL [69]	62.96	38.25	15.53	-	-	-	-	-	-
ZS	Luo *et al.* [37]	-	-	-	40.27	16.27	-	45.04	21.44	-
	TFVTG [68]	65.07	49.24	23.05	43.84	18.68	40.19	56.26	28.49	46.90
	TAG [26]	67.94	50.19	27.14	43.55	21.30	41.95	52.37	32.66	47.86
	Ours	**70.41**	**53.75**	**30.45**	**46.45**	**24.11**	**45.30**	**55.61**	**35.43**	**51.09**

routing and TMP remain active. Notably, this variant only marginally outperforms Fast only (44.70 vs. 44.34 mIoU), indicating that routing and structured reasoning provide little benefit when the underlying keyframes are uninformative, and that high-quality keyframe selection is a prerequisite for effective reasoning. Second, replacing TMP with a generic chain-of-thought prompt (w/o TMP) reduces mIoU by 2.80, confirming that task-specific structured reasoning outperforms unconstrained reasoning for temporal grounding. Third, fixing the keyframe budget to K=8 instead of adaptive stopping (w/o Adaptive K) drops mIoU by 1.67, showing that adapting the number of keyframes to each query is beneficial. Finally, comparing the two routing extremes, Fast only (44.34), which skips reasoning entirely, vs. Slow only (47.53), which applies TMP to all queries without spectral routing, reveals that applying TMP uniformly is much better than skipping it entirely, yet the full adaptive routing (48.93) surpasses

Table 4: Cross-dataset: train on ActivityNet Captions, test on Charades-STA.

Method	R@1 R@0.5	R@1 R@0.7	R@5 R@0.5	R@5 R@0.7
SCDM [60]	15.91	6.19	54.04	30.39
2D-TAN [64]	15.81	6.30	59.06	31.53
Debias-TLL [1]	21.45	10.38	62.34	32.90
TFVTG [68]	49.97	24.32	83.50	42.20
Ours	**53.17**	**27.84**	**87.43**	**46.97**

Table 5: Keyframe selection comparison on Charades-STA (IID).

Method	R@0.3	R@0.5	R@0.7	mIoU
K-medoids	67.24	47.53	26.18	45.73
Uniform + re-rank	68.15	48.62	27.08	46.52
Top-k + NMS	69.07	49.81	27.92	47.38
DPP (Ours)	**70.98**	**52.04**	**29.45**	**48.93**

Table 6: Component ablation on Charades-STA (IID). Each row removes one component from the full DART.

Variant	R@0.3	R@0.5	R@0.7	mIoU
Ours	**70.98**	**52.04**	**29.45**	**48.93**
w/o DPP	66.51	46.79	25.55	44.70
w/o Adaptive K	68.62	49.01	26.73	47.26
w/o TMP	68.14	48.65	25.92	46.13
Fast only	66.35	46.71	25.34	44.34
Slow only	68.71	49.80	26.78	47.53

Table 7: Inference efficiency on Charades-STA. Time per query on a single A100 GPU.

Method	Frames	Time (s)
TFVTG [68]	86	4.7
TAG [26]	86	3.4
DART (Slow)	12	5.1
DART (Adaptive)	12	3.9

both, confirming that selectively applying reasoning to complex queries while protecting simple ones from over-reasoning yields the best trade-off.

DPP vs. alternative keyframe selection. We compare our query-conditioned DPP against three alternatives in Tab. 5. All variants share the same downstream pipeline (spectral routing + dual-path localization) and differ only in the keyframe selection strategy. K-medoids clustering partitions frames into K clusters by visual similarity and selects each medoid, maximizing spatial coverage but entirely ignoring query relevance (45.73 mIoU). Uniform sampling with re-ranking first draws frames at equal intervals and then re-ranks them by query similarity to retain the top K; although query-aware, it is constrained by the quality of the initial uniformly-spaced candidate set (46.52 mIoU). Top-k similarity + NMS ranks all frames by query similarity, greedily selects the highest-scoring ones, and suppresses temporal neighbors within a fixed window; despite being the strongest baseline (47.38 mIoU), its greedy, independent selection still misses complementary event stages that a joint selection would capture. DPP jointly optimizes relevance and diversity through the query-conditioned kernel (Sec. 3.2), ensuring that selected keyframes are both highly relevant to the query and temporally diverse, yielding +1.55 mIoU over the best alternative.

4.4 Analysis

Spectral entropy reflects grounding difficulty. To validate that H_{spectral} is a reliable difficulty indicator, we randomly sample 1,000 queries from the ActivityNet Captions val set (200 per bin across five equal-width bins over [0, 1]) and evaluate all queries using the *Fast path only* (*i.e.*, pure feature matching

Table 8: Per-bin mIoU *vs.* H_{spectral} on 1,000 ActivityNet Captions val queries (200 per bin), evaluated using the Fast path only to ensure differences reflect query difficulty, not reasoning benefit. Higher entropy correlates with lower mIoU.

H_{spectral} bin	$[0, 0.2)$	$[0.2, 0.4)$	$[0.4, 0.6)$	$[0.6, 0.8)$	$[0.8, 1.0]$
mIoU (%)	49.2	43.7	37.1	31.8	27.4

without structured reasoning), so that performance differences purely reflect query difficulty rather than the benefit of structured reasoning. As shown in Tab. 8, mIoU decreases monotonically from 49.2% in the lowest bin to 27.4% in the highest, confirming that higher spectral entropy corresponds to objectively harder grounding instances. The sharp performance drop in high-entropy bins demonstrates that these queries cannot be resolved by feature matching alone and genuinely require the structured reasoning provided by the Slow path, validating H_{spectral} as an effective difficulty indicator for routing.

Efficiency comparison. Tab. 7 compares the inference cost of DART against zero-shot baselines on Charades-STA. Despite processing far fewer frames per query (12 vs. 86), DART with adaptive routing achieves 3.9s per query, which is faster than TFVTG (4.7s) and comparable to TAG (3.4s), while substantially outperforming both in accuracy (Tab. 1). The efficiency gain comes from two sources: (1) the DPP-based keyframe selection reduces the number of frames the LVLM must process, and (2) adaptive routing directs simple queries to the Fast path, avoiding the multi-step reasoning overhead of TMP. Compared to the Slow-only variant that applies TMP to every query, adaptive routing reduces inference time by 23.5% (5.1s $\rightarrow$ 3.9s) while also improving mIoU (47.53 $\rightarrow$ 48.93, Tab. 6), confirming that unnecessary reasoning on simple queries hurts both speed and accuracy.

5 Conclusion

We identified a reasoning gap in zero-shot video temporal grounding: feature-matching methods suffice for simple events but fail on complex queries requiring temporal and causal reasoning. We proposed DART to bridge this gap. Built around a query-conditioned DPP, DART selects diverse, query-relevant keyframes and derives a spectral entropy-based difficulty indicator from a single decomposition, routing simple queries to direct prediction and complex queries to our Temporal Markup Prompting for structured reasoning. Experiments on Charades-STA and ActivityNet Captions show that DART achieves state-of-the-art zero-shot performance under both IID and OOD settings, validating that deciding *when* to reason is as important as *how* to reason.

References

1. Bao, P., Mu, Y.: Learning sample importance for cross-scenario video temporal grounding. arXiv preprint arXiv:2201.02848 (2022)
2. Cai, S.: Iieu: Rethinking neural feature activation from decision-making. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 5796–5806 (October 2023)
3. Cai, S.: Adashift: Learning discriminative self-gated neural feature activation with an adaptive shift factor. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5947–5956 (June 2024)
4. Cai, S., Yuan, S., Chen, B., Mao, R., Wang, B.: Selection-as-nonlinearity: Bridging attention and activation via a joint game-decision lens for interpretable, discriminative visual representations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 11621–11631 (2026)
5. Cai, S., Zheng, S., Chen, B., Yuan, S., Xiao, C., Qin, J., WANG, B.: Toward principled flexible scaling for self-gated neural activation. In: International Conference on Learning Representations (ICLR) (2026)
6. Chen, B., Xu, Z., Kirmani, S., Ichter, B., Sadigh, D., Guibas, L., Xia, F.: SpatialVLM: Endowing vision-language models with spatial reasoning capabilities. In: CVPR. pp. 14455–14465 (2024)
7. Chen, J., Chen, X., Ma, L., Jie, Z., Chua, T.S.: Temporally grounding natural sentence in video. In: EMNLP. pp. 162–171 (2018)
8. Cheng, S., Zhang, J., Song, Q., Liu, S., Guo, Z., Zhang, X., Zhang, C., Li, X., Tu, Z.: Unison: Harmonizing motion, speech, and sound for human-centric audio-video generation (2026), https://arxiv.org/abs/2605.08729
9. Cheng, S., Zhang, J., Liu, Y., Xiao, A., Tu, Z.: Owlsight: A robust illumination adaptation framework for dark video human action recognition. IEEE Transactions on Circuits and Systems for Video Technology **36**(4), 4550–4564 (2025). https://doi.org/10.1109/TCSVT.2025.3632359
10. Duan, X., Huang, W., Gan, C., Wang, J., Zhu, W., Huang, J.: Weakly supervised dense event captioning in videos. NeurIPS **31** (2018)
11. Gao, J., Sun, C., Yang, Z., Nevatia, R.: TALL: Temporal activity localization via language query. In: ICCV. pp. 5267–5275 (2017)
12. Gao, J., Xu, C.: Learning video moment retrieval without a single annotated video. IEEE TCSVT **32**(3), 1646–1657 (2021)
13. Gao, M., Davis, L.S., Socher, R., Xiong, C.: WSLLN: Weakly supervised natural language localization networks. arXiv preprint arXiv:1909.00239 (2019)
14. Hu, K., Gao, F., Nie, X., Zhou, P., Tran, S., Neiman, T., Wang, L., Shah, M., Hamid, R., Yin, B., Chilimbi, T.M.: M-llm based video frame selection for efficient video understanding. 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 13702–13712 (2025), https://api.semanticscholar.org/CorpusID:276647361
15. Huang, B., Wang, X., Chen, H., Song, Z., Zhu, W.: VTimeLLM: Empower LLM to grasp video moments. In: CVPR. pp. 14271–14280 (2024)
16. Huang, J., Jin, H., Gong, S., Liu, Y.: Video activity localisation with uncertainties in temporal boundary. In: ECCV. pp. 724–740 (2022)
17. Huang, J., Liu, Y., Gong, S., Jin, H.: Cross-sentence temporal and semantic relations in video activity localisation. In: ICCV. pp. 7199–7208 (2021)
18. Huang, Y., Yang, L., Sato, Y.: Weakly supervised temporal sentence grounding with uncertainty-guided self-training. In: CVPR. pp. 18908–18918 (2023)

19. Jang, J., Park, J., Kim, J., Kwon, H., Sohn, K.: Knowing where to focus: Event-aware transformer for video grounding. In: ICCV. pp. 13846–13856 (2023)
20. Jeon, M., Ma, M., Kim, J.: Dbcon: Dual bias control in zero-shot video moment retrieval. IEEE Access **13**, 167439–167448 (2025), https://api.semanticscholar.org/CorpusID:281516316
21. Jeon, M., Yoon, S., Kim, J., Kim, J.: GranAlign: Granularity-aware alignment framework for zero-shot video moment retrieval. In: AAAI (2026)
22. Kim, D., Park, J., Lee, J., Park, S., Sohn, K.: Language-free training for zero-shot video grounding. In: WACV. pp. 2539–2548 (2023)
23. Krishna, R., Hata, K., Ren, F., Fei-Fei, L., Niebles, J.C.: Dense-captioning events in videos. In: ICCV (2017)
24. Kulesza, A., Taskar, B.: Determinantal point processes for machine learning. Foundations and Trends in Machine Learning **5**(2–3), 123–286 (2012)
25. Kuo, C.C., Glover, F., Dhir, K.S.: Analyzing and modeling the maximum diversity problem by zero-one programming. Decision Sciences **24**(6), 1171–1185 (1993). https://doi.org/10.1111/j.1540-5915.1993.tb00509.x
26. Lee, J.S., Lee, S., Ahn, J.C., Choi, Y., Lee, J.H.: Tag: A simple yet effective temporal-aware approach for zero-shot video temporal grounding. ArXiv **abs/2508.07925** (2025), https://api.semanticscholar.org/CorpusID:280567060
27. Lei, J., Berg, T.L., Bansal, M.: Detecting moments and highlights in videos via natural language queries. In: NeurIPS. pp. 11846–11858 (2021)
28. Lei, J., Yu, L., Bansal, M., Berg, T.L.: TVQA: Localized, compositional video question answering. In: EMNLP. pp. 1369–1379 (2018)
29. Li, J., Xie, J., Qian, L., Zhu, L., Tang, S., Wu, F., Yang, Y., Zhuang, Y., Wang, X.E.: Compositional temporal grounding with structured variational cross-graph correspondence learning. In: CVPR. pp. 3032–3041 (2022)
30. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: International conference on machine learning. pp. 19730–19742. PMLR (2023)
31. Li, K., He, Y., Wang, Y., Li, Y., Wang, W., Luo, P., Wang, Y., Wang, L., Qiao, Y.: Videochat: chat-centric video understanding. Science China Information Sciences **68** (2023), https://api.semanticscholar.org/CorpusID:258588306
32. Li, Z., Xu, Q., Zhang, D., Song, H., Cai, Y., Qi, Q., Zhou, R., Pan, J., Li, Z., Vu, V.T., et al.: GroundingGPT: Language enhanced multi-modal grounding model. arXiv preprint arXiv:2401.06071 (2024)
33. Liu, D., Qu, X., Di, X., Cheng, Y., Xu, Z., Zhou, P.: Memory-guided semantic learning network for temporal sentence grounding. arXiv preprint arXiv:2201.00454 (2022)
34. Liu, H., Li, C., Li, Y., Li, B., Zhang, Y., Shen, S., Lee, Y.J.: LLaVA-NeXT: Improved reasoning, OCR, and world knowledge (January 2024), https://llava-vl.github.io/blog/2024-01-30-llava-next/
35. Liu, R., Geng, J., Wu, A.J., Sucholutsky, I., Lombrozo, T., Griffiths, T.L.: Mind your step (by step): Chain-of-thought can reduce performance on tasks where thinking makes humans worse. In: ICML (2025)
36. Luo, D., Huang, J., Gong, S., Jin, H., Liu, Y.: Towards generalisable video moment retrieval: Visual-dynamic injection to image-text pre-training. In: CVPR. pp. 23045–23055 (2023)
37. Luo, D., Huang, J., Gong, S., Jin, H., Liu, Y.: Zero-shot video moment retrieval from frozen vision-language models. 2024 IEEE/CVF Winter Conference

on Applications of Computer Vision (WACV) pp. 5452–5461 (2023), https://api.semanticscholar.org/CorpusID:261531052
38. Maaz, M., Rasheed, H.A., Khan, S.H., Khan, F.S.: Video-chatgpt: Towards detailed video understanding via large vision and language models. In: Annual Meeting of the Association for Computational Linguistics (2023), https://api.semanticscholar.org/CorpusID:259108333
39. Mithun, N.C., Paul, S., Roy-Chowdhury, A.K.: Weakly supervised video moment retrieval from text queries. In: CVPR. pp. 11592–11601 (2019)
40. Mun, J., Cho, M., Han, B.: Local-global video-text interactions for temporal grounding. In: CVPR. pp. 10810–10819 (2020)
41. Nam, J., Ahn, D., Kang, D., Ha, S.J., Choi, J.: Zero-shot natural language video localization. In: ICCV. pp. 1470–1479 (2021)
42. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. ArXiv **abs/2103.00020** (2021), https://api.semanticscholar.org/CorpusID:231591445
43. Roy, O., Vetterli, M.: The effective rank: A measure of effective dimensionality. 2007 15th European Signal Processing Conference pp. 606–610 (2007), https://api.semanticscholar.org/CorpusID:12184201
44. Sigurdsson, G.A., Varol, G., Wang, X., Farhadi, A., Laptev, I., Gupta, A.: Hollywood in homes: Crowdsourcing data collection for activity understanding. In: ECCV. pp. 510–526 (2016)
45. Song, Q., He, Y., Zhang, Y., Cheng, S., He, Z., Guo, Z., Zhang, C., Li, X., Jiang, C.: Interactiveavatar: Real-time streaming video generation for consistent and intent-aware avatars (2026), https://arxiv.org/abs/2606.22905
46. Tu, Z., Zhang, Z., Gong, J., Yuan, J., Du, B.: Informative sample selection model for skeleton-based action recognition with limited training samples. IEEE Transactions on Image Processing **34**, 7335–7346 (2025)
47. Wang, G., Wu, X., Liu, Z., Yan, J.: Prompt-based zero-shot video moment retrieval. In: ACM MM. pp. 413–421 (2022)
48. Wang, Y., Xu, B., Yue, Z., Xiao, Z., Wang, Z., Zhang, L., Yang, D., Wang, W., Jin, Q.: Timezero: Temporal video grounding with reasoning-guided lvlm. ArXiv **abs/2503.13377** (2025), https://api.semanticscholar.org/CorpusID:281707035
49. Wang, Y., Li, K., Li, X., Yu, J., He, Y., Chen, G., Pei, B., Zheng, R., Xu, J., Wang, Z., et al.: InternVideo2: Scaling foundation models for multimodal video understanding. In: ECCV (2024)
50. Wang, Z., Wang, L., Wu, T., Li, T., Wu, G.: Negative sample matters: A renaissance of metric learning for temporal grounding. In: AAAI. pp. 2613–2621 (2022)
51. Wattasseril, J.I., Shekhar, S., Döllner, J., Trapp, M.: Zero-shot video moment retrieval using BLIP-based models. In: International Symposium on Visual Computing. pp. 160–171 (2023)
52. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Chi, E.H., Xia, F., Le, Q., Zhou, D.: Chain of thought prompting elicits reasoning in large language models. ArXiv **abs/2201.11903** (2022), https://api.semanticscholar.org/CorpusID:246411621
53. Wu, J., Li, G., Liu, S., Lin, L.: Tree-structured policy based progressive reinforcement learning for temporally language grounding in video. In: AAAI. vol. 34, pp. 12386–12393 (2020)

54. Xiao, A., Cheng, S., Xu, Y., Ren, Y., Chen, H., Yokoya, N.: Geommbench and geommagent: Toward expert-level multimodal intelligence in geoscience and remote sensing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 34843–34853 (June 2026)
55. Xu, Y., Sun, Y., Xie, Z., Zhai, B., Du, S.: Vtg-gpt: Tuning-free zero-shot video temporal grounding with gpt. ArXiv **abs/2403.02076** (2024), https://api.semanticscholar.org/CorpusID:268035181
56. Xu, Y., Sun, Y., Zhai, B., Li, M., Liang, W., Du, S.: Zero-shot video moment retrieval via off-the-shelf multimodal large language models. AAAI pp. 8978–8986 (2025)
57. Yang, L., Kong, Q., Yang, H.K., Kehl, W., Sato, Y., Kobori, N.: Deco: Decomposition and reconstruction for compositional temporal grounding via coarse-to-fine contrastive ranking. In: CVPR. pp. 23130–23140 (2023)
58. Yang, X., Feng, F., Ji, W., Wang, M., Chua, T.S.: Deconfounded video moment retrieval with causal intervention. In: SIGIR. pp. 1–10 (2021)
59. Yuan, Y., Lan, X., Wang, X., Chen, L., Wang, Z., Zhu, W.: A closer look at temporal sentence grounding in videos: Dataset and metric. In: ACM Workshop on Human-Centric Multimedia Analysis. pp. 13–21 (2021)
60. Yuan, Y., Ma, L., Wang, J., Liu, W., Zhu, W.: Semantic conditioned dynamic modulation for temporal sentence grounding in videos. NeurIPS **32** (2019)
61. Zhang, H., Li, X., Bing, L.: Video-llama: An instruction-tuned audio-visual language model for video understanding. In: Conference on Empirical Methods in Natural Language Processing (2023), https://api.semanticscholar.org/CorpusID:259075356
62. Zhang, H., Sun, A., Jing, W., Zhou, J.T.: Span-based localizing network for natural language video localization. In: ACL. pp. 6543–6554 (2020)
63. Zhang, J., Guo, Y., Potamias, R.A., Deng, J., Xu, H., Ma, C.: VTimeCoT: Thinking by drawing for video temporal grounding and reasoning. In: ICCV. pp. 24203–24213 (2025)
64. Zhang, S., Peng, H., Fu, J., Luo, J.: Learning 2d temporal adjacent networks for moment localization with natural language. In: AAAI. vol. 34, pp. 12870–12877 (2020)
65. Zhang, Z., Tu, Z., Yuan, J., Soh, D.W., Du, B.: Leveraging text-to-image diffusion models for unsupervised visual object tracking. IEEE Transactions on Pattern Analysis and Machine Intelligence (2026)
66. Zhang, Z., Xu, L., Peng, D., Rahmani, H., Liu, J.: Diff-tracker: text-to-image diffusion models are unsupervised trackers. In: European Conference on Computer Vision. pp. 319–337. Springer (2024)
67. Zhang, Z., Zhou, Y., Peng, D., Lim, J.H., Tu, Z., Soh, D.W., Foo, L.G.: Visual prompting for one-shot controllable video editing without inversion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7784–7794 (2025)
68. Zheng, M., Cai, X., Chen, Q., Peng, Y., Liu, Y.: Training-free video temporal grounding using large-scale pre-trained models. In: European Conference on Computer Vision (2024), https://api.semanticscholar.org/CorpusID:272146312
69. Zheng, M., Gong, S., Jin, H., Peng, Y., Liu, Y.: Generating structured pseudo labels for noise-resistant zero-shot video sentence localization. In: ACL. pp. 14197–14209 (2023)
70. Zheng, M., Huang, Y., Chen, Q., Liu, Y.: Weakly supervised video moment localization with contrastive negative sample mining. In: AAAI (2022)

71. Zheng, M., Huang, Y., Chen, Q., Peng, Y., Liu, Y.: Weakly supervised temporal sentence grounding with gaussian-based contrastive proposal learning. In: CVPR (2022)