SpaceVLLM: Endowing Multimodal Large Language Model with Spatio-Temporal Video Grounding Capability

Jiankang Wang[1*], Zhihan Zhang[1*], Zhihang Liu[1], Yang Li[2], Jiannan Ge[1], Hongtao Xie[1†], Yongdong Zhang[1]

[1]University of Science and Technology of China
[2]Renmin University of China
{wangjiankang, zhangzhihan, liuzhihang, gejn}@mail.ustc.edu.cn, liyang03@ruc.edu.cn, {htxie, zhyd73}@ustc.edu.cn

Abstract

Multimodal Large Language Models (MLLMs) have shown remarkable progress in temporal or spatial localization tasks, but struggle with joint spatio-temporal video grounding (STVG). We identify two key bottlenecks hindering this capability: (1) the sheer number of visual tokens makes long-range and fine-grained visual modeling challenging; (2) generating a long sequence of bounding boxes in text makes it hard to accurately align each box with its specific video frame. Distinct from prior efforts that rely on attaching complex modules, we argue for a more elegant paradigm that unlocks the inherent potential of MLLMs and leverages their strengths. To this end, we propose ***SpaceVLLM***, a MLLM equipped with spatio-temporal video grounding capabilities. Specifically, we propose Spatio-Temporal Aware Queries, interleaved with video frames, to guide the MLLM in capturing both static appearance and dynamic motion features. We further present a lightweight Query-Guided Space Head that maps queries to precise spatial coordinates, bypassing the need for direct textual coordinate generation and enabling the MLLM to focus on video understanding. To further facilitate research in this area, we propose an automated data synthesis pipeline to construct **V-STG** dataset, comprising 110K STVG instances. Extensive experiments show that *SpaceVLLM* achieves the state-of-the-art performance on STVG benchmarks and maintains strong performance on various video understanding tasks, validating our approach's effectiveness.

Code — https://github.com/Jayce1kk/SpaceVLLM
Extended version — https://arxiv.org/abs/2503.13983

Introduction

Multimodal Large Language Models (MLLMs) have recently demonstrated significant advancements in multimodal understanding (Li et al. 2023; Lin et al. 2023a). With the rapid development of MLLMs, an increasing number of works focus on multimodal comprehension from either a temporal or spatial perspective. Some studies (Guo et al. 2024a,b) aim to enhance the ability of MLLMs to perceive temporal information in tasks such as video temporal grounding, while others (Chen et al. 2023; Ma et al. 2024) concentrate on localizing referred objects within a single image. However, spatio-temporal video grounding (STVG) (Zhang et al. 2021), an important capability for multimodal video understanding that requires simultaneously localizing targets both spatially and temporally in untrimmed videos given textual descriptions, remains largely unexplored in existing MLLMs.

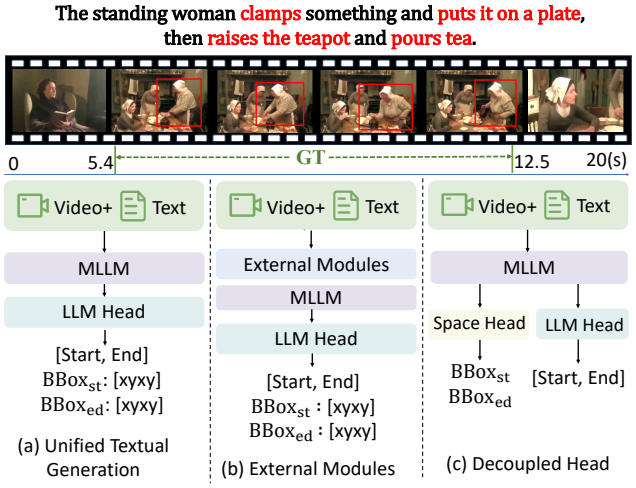

Figure 1: Three methods for spatio-temporal video grounding using multimodal large language model.

A straightforward approach to Spatio-Temporal Video Grounding (STVG) is to treat it as a unified text generation task, fine-tuning powerful MLLMs (Zhang et al. 2024; Bai et al. 2025) to output a string of timestamps and bounding boxes (Figure 1(a)). However, since the model is required to generate frame-wise box information in textual form for a large number of video frames, this approach poses significantly greater challenges compared to generating a single box for a still image. There are two main bottlenecks: (1) Long-range and fine-grained visual modeling barrier: The substantial visual tokens from video frames overwhelm the MLLM, making it difficult to maintain both long-range context and precise focus on fine-grained details necessary for accurate grounding. (2) Alignment ambiguity: The model

*These authors contributed equally.
†Corresponding author
Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

generates a long sequence of coordinates, making it difficult to ensure that each box is accurately aligned with its corresponding video frame. As a result, MLLMs are not adept at generating long sequences of high-precision, structured coordinates as raw text. Although recent works (Li et al. 2025) have partially alleviated these issues by introducing complex, pre-trained modules for spatio-temporal feature extraction and alignment (see Figure 1(b)), we take a different view. We argue that MLLMs, after large-scale video training, already possess strong spatio-temporal modeling capacity, but currently lack efficient guidance to fully unlock this potential. Therefore, we advocate for a more elegant paradigm: *unlocking the inherent potential of MLLMs and leveraging their strengths.*

To realize this new paradigm, we introduce ***SpaceVLLM***, a simple and efficient Decoupled Head architecture that lets the MLLM focus on its strengths in high-level video understanding, while offloading specialized tasks to a dedicated head (Figure 1(c)). Specifically, to effectively capture the fine-grained spatio-temporal details, we introduce Spatio-Temporal Aware Queries. These are parameter-free special tokens interleaved with the frame features, guiding the model to perform short-range modeling for each frame rather than relying on long-range dependencies. They also guide the MLLM to extract static appearance and dynamic motion features from the video input. To generate faithful coordinates, we propose a lightweight Query-Guided Space Head. This head takes the semantically rich queries output from the MLLM and maps them into coordinates, freeing the MLLM from the burden of spatial coordinate generation. Further progress of MLLMs in STVG is also constrained by the limited availability of suitable training datasets. Most existing datasets are designed to address either temporal grounding or spatial grounding in isolation. However, STVG datasets are extremely limited and typically require expensive manual annotation. To address this, we propose an automated data synthesis pipeline to construct a high-quality Video Spatio-Temporal Grounding (**V-STG**) comprising 110K instances. V-STG features a wide range of video sources and diverse localization targets, aiming to advance multimodal spatio-temporal understanding. We conduct extensive experiments on 11 benchmark datasets. The results demonstrate that our model achieves state-of-the-art performance on Spatio-Temporal Video Grounding, while also maintaining strong results on various video understanding tasks, highlighting the effectiveness of our approach.

Our contributions can be summarized as follows:

- We propose a new paradigm for STVG that unlocks the latent potential of MLLMs, rather than relying on complex external modules. Based on the paradigm, we present *SpaceVLLM*, a MLLM equipped with spatio-temporal video grounding capability.
- We propose a high-quality Video Spatio-Temporal Grounding dataset (V-STG) with 110K instances, facilitating fine-grained spatial-temporal understanding.
- We conduct experiments on 11 benchmarks and achieve state-of-the-art performance on multiple benchmarks, demonstrating the effectiveness of our approach.

Related Work

Spatio-Temporal Video Grounding

Spatio-temporal video grounding (STVG) is a critical multimodal task aiming to localize a text-described target in both space and time within an untrimmed video (Zhang et al. 2021). Traditional STVG research evolved from two-stage methods reliant on pre-trained detectors (Tang et al. 2021; Zhang et al. 2021) to more streamlined one-stage approaches that directly predict spatio-temporal proposals (Yang et al. 2022; Gu et al. 2024). Despite this progress, STVG remains largely unexplored in MLLMs. The recent work LLaVA-ST (Li et al. 2025), which proposes LAPE for coordinate alignment and introduces STP to preserve spatio-temporal information. In this work, we explore a more elegant and effective paradigm to unlock the inherent potential of MLLMs for accomplishing Spatio-Temporal Video Grounding.

Multimodal Large Language Models

Recently, Multimodal Large Language Models (MLLMs) have made significant advances in video understanding. Existing video LLMs have achieved strong results in visual question answering, video captioning, and reasoning(Li et al. 2023; Zhang, Li, and Bing 2023; Maaz et al. 2024b). Some works focus on temporal localization, such as VTimeLLM (Huang et al. 2024), which adopts a boundary-aware training strategy, and TRACE(Guo et al. 2024b), which introduces causal event modeling. Others address spatial localization, as in Groma (Ma et al. 2024) with localized visual tokenization, and GroundingGPT (Li et al. 2024b), which enhances multimodal grounding with language cues. However, current models still struggle with joint spatial-temporal grounding. In this paper, we propose *SpaceVLLM*, a novel model that empowers the MLLM with spatio-temporal video grounding capability.

Method

Overview We first introduce the Spatio-Temporal Aware Query. Next, we present the Query-Guided Space Head and two alternative decoder architectures. Finally, we describe the training objectives.

Spatio-Temporal Aware Query The Spatio-Temporal Aware Query (STA query) is designed as a parameter-free special token whose function is not to learn representations independently, but rather to guide the powerful MLLM to summarize the spatio-temporal information of each frame and embed that representation into the query itself.

Specifically, we first sample N_v frames from each video, denoted as $\mathbf{X} = \{x_i\}_{i=1}^{N_v}$, and associate each frame with a special token $\{<q_i>\}_{i=1}^{N_v}$, which serve as STA queries. The vision encoder extracts visual features from each frame, which are then projected into the textual embedding space, resulting in a sequence of visual embeddings $\mathbf{V} \in \mathbb{R}^{N_v \times S \times D}$, where S is the number of visual tokens per frame and D is the hidden dimension. In parallel, the text tokenizer converts each special token into a text embedding, yielding $\mathbf{Q} \in \mathbb{R}^{N_v \times 1 \times D}$. Finally, we construct the extended visual representation $\mathbf{H}$ by interleaving and concatenating

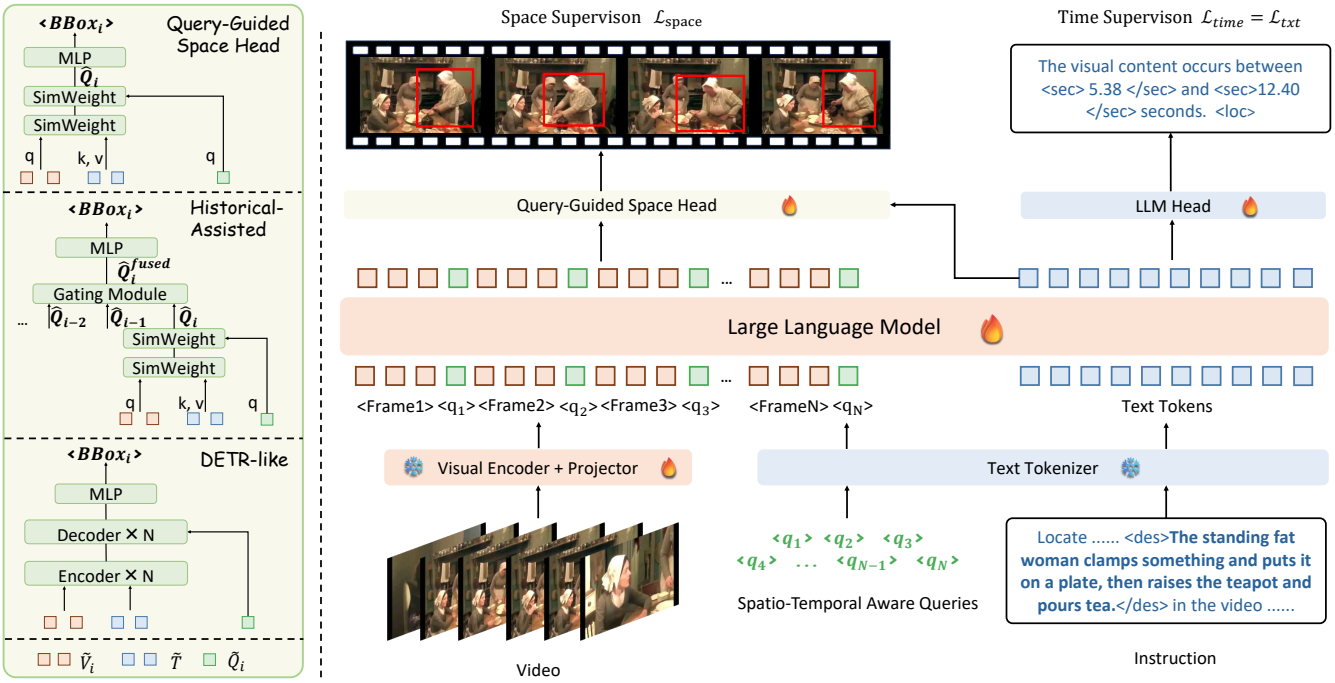

Figure 2: The Overall Architecture of *SpaceVLLM*. In *SpaceVLLM*, A set of ordered Spatio-Temporal Aware Queries is interleaved with visual tokens of each video frame to capture spatio-temporal information. The LLM's last-layer query embeddings, combined with corresponding visual and description embeddings, are fed into the Query-Guided Space Head to predict frame-wise coordinates. The left figure illustrates three types of spatial decoder architectures.

the visual and STA query embeddings:

$$\mathbf{H} = \bigoplus_{i=0}^{N_v-1} (\mathbf{V}_i \oplus \mathbf{Q}_i), \quad \mathbf{H} \in \mathbb{R}^{N_v \times (S+1) \times D}, \tag{1}$$

where $\mathbf{V}_i$ denotes the visual features of the i-th frame, $\mathbf{Q}_i$ represents the i-th STA query embedding, and $\oplus$ indicates row-wise concatenation. Subsequently, $\mathbf{H}$, together with the user instruction, is fed into the LLM. We extract the last-layer hidden embeddings from the LLM to obtain the Spatio-Temporal Aware Queries $\tilde{\mathbf{Q}} \in \mathbb{R}^{N_v \times 1 \times D}$, the visual features $\tilde{\mathbf{V}} \in \mathbb{R}^{N_v \times S \times D}$, and the textual features $\tilde{\mathbf{T}} \in \mathbb{R}^{1 \times L \times D}$, where L denotes the length of the given caption.

Query-Guided Space Head We propose a novel decoupled head architecture in *SpaceVLLM*, illustrated in Figure 2, designed to free the MLLM from the cumbersome and imprecise task of generating coordinate sequences. In this architecture, tasks are divided: the LLM head focuses on the output of multimodal understanding, while an equally lightweight Query-Guided Space Head specializes in mapping the STA queries into frame-specific coordinates. The core operation is a standard, **parameter-free** cross-attention that functions as a feature similarity calculation. For brevity and clarity throughout the paper, we will refer to it as **Similarity Weighting** in the subsequent text, and denote it as `SimWeight(·)` in Figure 2 and formulas.

Specifically, for the i-th frame, similarity weighting is first performed between the visual embeddings $\tilde{\mathbf{V}}_i \in \mathbb{R}^{1 \times S \times D}$ and the textual embeddings of the caption $\tilde{\mathbf{T}} \in \mathbb{R}^{1 \times L \times D}$ to obtain text-enhanced visual features $\hat{\mathbf{V}}_i$. Next, the spatio-temporal aware query $\tilde{\mathbf{Q}}_i \in \mathbb{R}^{1 \times 1 \times D}$ further interacts with these features through similarity weighting, resulting in the target-enhanced query $\hat{\mathbf{Q}}_i \in \mathbb{R}^{1 \times 1 \times D}$. This process can be formulated as follows:

$$\begin{aligned} \hat{\mathbf{V}}_i &= \mathrm{SimWeight}(\tilde{\mathbf{V}}_i, \tilde{\mathbf{T}}, \tilde{\mathbf{T}}), \\ \hat{\mathbf{Q}}_i &= \mathrm{SimWeight}(\tilde{\mathbf{Q}}_i, \hat{\mathbf{V}}_i, \hat{\mathbf{V}}_i). \end{aligned} \tag{2}$$

Finally, $\hat{\mathbf{Q}}_i$ is fed into a lightweight Multi-Layer Perceptron (MLP) to obtain the spatial bounding box coordinates $\hat{\mathbf{B}}_i$:

$$\hat{\mathbf{B}}_i = \mathrm{MLP}(\hat{\mathbf{Q}}_i), \quad \hat{\mathbf{B}}_i \in [0,1]^4, \tag{3}$$

where $\hat{\mathbf{B}}_i = (c_x, c_y, w, h)$ represents the normalized center coordinates and size of the bounding box for the i-th frame.

Alternative Decoder Architectures To validate our core design principle that MLLMs possess strong inherent spatio-temporal modeling capacity without needing complex external modules, we systematically explore two alternative decoder architectures, as shown in Figure 2.

(1) **Historical-Assisted Decoder**. To test if the MLLM inherently possesses the ability to capture dynamic motion features—a key capability for STVG—we design this decoder. This decoder builds upon our proposed Query-Guided Space Head by incorporating a gating module that fuses the current frame's target-enhanced query with queries from

historical frames. Specifically, for the i-th frame, the fused query $\hat{\mathbf{Q}}_i^{\text{fused}}$ is obtained using a gated mixture-of-experts (MoE) module, and is computed as:

$$\hat{\mathbf{Q}}_i^{\text{fused}} = \sum_{k=i-n}^{i} \alpha_{i,k} \cdot \mathbf{h}_k(\hat{\mathbf{Q}}_k), \quad \hat{\mathbf{Q}}_i^{\text{fused}} \in \mathbb{R}^{1\times 1\times D}, \quad (4)$$

where $\hat{\mathbf{Q}}_k$ denotes the target-enhanced query of the k-th frame, and n is a hyperparameter controlling the range of historical frames to be fused (by default, $n = 1$). $\mathbf{h}_k(\cdot)$ is the expert function applied to each query; it is the identity mapping if $k = i$, and typically an MLP otherwise. The gating weights $\alpha_{i,k}$ are normalized and dynamically computed by a gating network $G(\cdot)$ (e.g., an MLP). For example, when $n = 1$, the gating weights are computed as:

$$[\alpha_{i,i-1}, \alpha_{i,i}] \;=\; \text{softmax}\Big(G\big([\hat{\mathbf{Q}}_{i-1}, \hat{\mathbf{Q}}_i]\big)\Big), \quad (5)$$

where $[\cdot]$ denotes vector concatenation. Finally, the coordinates $\hat{\mathbf{B}}_i$ are obtained via $\text{MLP}(\hat{\mathbf{Q}}_i^{\text{fused}})$. We record all gating weights $\alpha_{i,k}$ of $\hat{\mathbf{Q}}_k$ during training and analyze whether the model needs historical frames.

(2) **DETR-like Architecture**. We implement this architecture, an approach inspired by traditional methods in spatial localization, to validate whether a more complex structure yields better performance. In detail, visual and textual embeddings are first fused by a multi-layer Transformer encoder. These joint representations are then processed by a multi-layer Transformer decoder, where they interact with the STA queries. Finally, an MLP head uses the decoder's output to predict the bounding box coordinates.

In the ablation study section, we systematically analyze three architectures.

Training Objectives For every spatio-temporal training sample, we decompose the loss into two components: the time loss $\mathcal{L}_{time}$ and the space loss $\mathcal{L}_{space}$. Each sample has a ground-truth bounding box sequence $\mathbf{B} = \{b_t\}_{t=t_s}^{t_e}$ and the corresponding text containing the start and end timestamps $\mathbf{y}_{txt}$. For spatial localization, we involve the box prediction loss $\mathcal{L}_{space}$ with loss weights λ_{L_1} and λ_{giou} as follows:

$$\mathcal{L}_{space} = \lambda_{L_1}\mathcal{L}_{L_1}(\hat{\mathbf{B}}, \mathbf{B}) + \lambda_{giou}\mathcal{L}_{giou}(\hat{\mathbf{B}}, \mathbf{B}), \quad (6)$$

where $\mathcal{L}_{L_1}$ and $\mathcal{L}_{\text{giou}}$ are the L_1 loss and generalized IoU loss (Rezatofighi et al. 2019) on the bounding boxes respectively. Note that $\mathcal{L}_{space}$ only considers predictions in $[t_s, t_e]$. As for temporal localization, we leverage MLLM to predict the time range. The time loss is computed using the auto-regressive cross-entropy loss $\mathcal{L}_{txt}$ for text generation. Given the ground-truth targets $\mathbf{y}_{txt}$, $\mathcal{L}_{time}$ can be denoted as $\mathcal{L}_{time} = \mathcal{L}_{txt}(\hat{\mathbf{y}}_{txt}, \mathbf{y}_{txt})$, where $\hat{\mathbf{y}}_{txt}$ refers the LLM's text output. The overall objective $\mathcal{L}$ is the weighted sum of these losses, determined by λ_{time} and λ_{space}:

$$\mathcal{L} = \lambda_{time}\mathcal{L}_{time} + \lambda_{space}\mathcal{L}_{space}. \quad (7)$$

During inference for the spatio-temporal video grounding task, the LLM head outputs both the temporal boundaries and a special signal token "<loc>". Upon recognizing the "<loc>" token, the STA queries within the predicted temporal range are selected and decoded to generate the bounding boxes. During training, the temporal boundaries are provided by the ground-truth annotations.

Data Curation and Training Datasets

Data Curation

Existing spatio-temporal video grounding (STVG) datasets (Tang et al. 2021; Zhang et al. 2020) are limited in scale and rely heavily on manual annotation. To facilitate efficient training of MLLMs, we propose an effective pipeline for synthesizing high-quality STVG data. Figure 3 shows the pipeline of data synthesis, which contains four components: (1) Analyzer for object extraction. (2) Annotator for box generation. (3) Refiner for time boundary. (4) Filter for bounding box.

(1) Analyzer for Object Extraction. We first collect a wide range of video data from Charades-STA (Gao et al. 2017), TACoS (Regneri et al. 2013), DiDeMo (Hendricks et al. 2017), and Intervid (Wang et al. 2023), each paired with captions and relevant timestamps. We then use Qwen2.5-72B (Yang et al. 2024b) as an analyzer to extract localizable objects from the captions, such as people, animals, objects, vehicles, etc. In subsequent processes, we prioritize locating the subject of the caption, followed by other mentioned objects. Figure 4 presents the distribution of video sources and target object categories.

(2) Annotator for Box Generation. We annotate the bounding box for each frame within the timestamps range. To ensure precise spatial localization, we employ Grounding-DINO (Liu et al. 2024a) to extract bounding boxes using the identified object as a text prompt. It allows us to generate multiple high-confidence bounding boxes for each frame, filtering out those with confidence scores below 0.3.

(3) Refiner for Time Boundary. The timestamp annotations in video datasets are not always precise. For instance, DiDeMo (Hendricks et al. 2017) adopts a time interval that is an integer multiple of five, leading to many start and end times that do not correspond to any actual objects. To address this issue, we refine the temporal boundaries by adjusting timestamps to better align with actual object appearances based on the Grounding-DINO's output. Additionally, we filter out adjusted timestamps that are either shorter than 2 seconds or longer than 120 seconds.

(4) Filter for Bounding Box. Obtaining high-quality bounding boxes for each frame is crucial for precise spatial-temporal localization. To refine our annotations, we implement a multi-step filtering process. First, filter the inaccurate instances. If a frame contains more than three bounding boxes, we consider it a case of either multiple objects or inaccurate localization and thus discard it. For the remaining frames, we retain only the bounding box with the highest confidence score. Second, filter through the object size. Object sizes should not fluctuate drastically between consecutive frames. To enforce this, we compare the bounding box area of each frame with that of the bounding box across adjacent frames. We remove samples where the box area is less than half or more than twice the reference box's area, ensuring stable object localization. Through this filtering pipeline, we eliminate approximately 40% of samples.

For HCSTVG and VidSTG, some spatio-temporal annotations are inaccurate. To address this, we review each instance using the last three steps of our pipeline. Instances with significant discrepancies are either removed or corrected. Ul-

Stage	Task	Datasets	Samples
Multi-Task Instruction Tuning	Video Temporal Grounding	VTG-IT (Guo et al. 2025), Charades-STA (Gao et al. 2017), HiREST (Zala et al. 2023)	50K
	Spatio-Temporal Video Grounding	V-STG	110K
	Referring Expression Comprehension	RefCOCO, RefCOCO+ (Kazemzadeh et al. 2014), RefCOCOg (Mao et al. 2016)	320K
	Video Question Answering	NeXTQA (Xiao et al. 2021), ActivityNetQA (Yu et al. 2019), CLEVRER (Yi et al. 2020), STAR (Wu et al. 2024), PerceptionTest (Patraucean et al. 2023)	100K
	Video Caption & Conversation	ShareGemini (Share 2024), ShareGPT4Video (Chen et al. 2024), VCG-Plus (Maaz et al. 2024a)	100K

Table 1: Overview of Datasets Used in Training for Various Tasks.

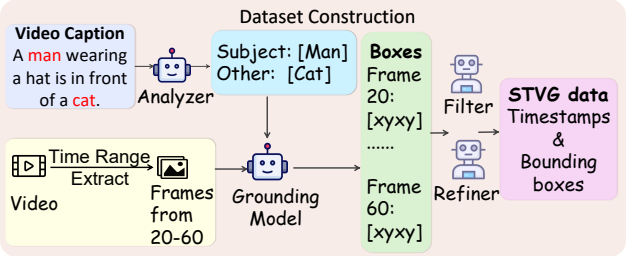

Figure 3: Pipeline of data synthesis for STVG task.

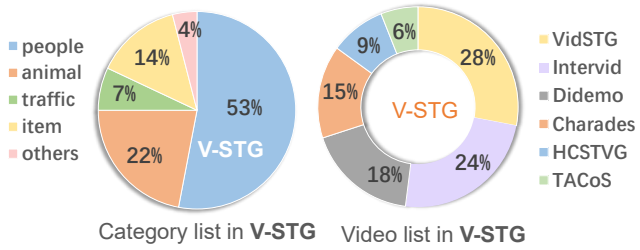

Figure 4: Data characteristics of V-STG for STVG task.

timately, these curated datasets, together with our synthesized data, comprise a high-quality Video Spatio-Temporal Grounding (**V-STG**) dataset containing 110K samples.

Training Datasets

As shown in Table 1, in addition to Spatio-Temporal Video Grounding, we also collect data for Video Temporal Grounding, Referring Expression Comprehension, Video Question Answering, Video Captioning, and Video Conversation in the training set, forming a unified multi-task video dataset, called **UniViT**. This enables us to perform multi-task instruction tuning and further investigate the impact of our approach on other related tasks.

Experiments

Implementation Details

We employ SigLIP (Zhai et al. 2023) as the vision encoder and Qwen2 (Yang et al. 2024a) as the LLM. The Query-Guided Space Head is adopted in all main experiments. We use AdamW (Loshchilov and Hutter 2017) optimizer with the learning rate and weight decay set to 1e-5 and 0, respectively. We also adopt cosine as the learning rate scheduler, where the warmup ratio is set to 0.03. We train the *SpaceVLLM* with 16 NVIDIA H100 GPUs in 24 hours based on the LLaVA-Video (Zhang et al. 2024) model and sample 64 frames per video. The loss weights are set as follows: $\lambda_{time} = 1.0$, $\lambda_{space} = 1.0$, $\lambda_{\mathcal{L}_1} = 3.0$, and $\lambda_{\text{giou}} = 1.0$.

Main Results

For a comprehensive evaluation, we consider 11 benchmarks that cover Spatio-Temporal Video Grounding (STVG), Referring Expression Comprehension (REC), Video Temporal Grounding (VTG), and video understanding tasks. Notably, all our results across different benchmarks are obtained using the same model by modifying the prompts.

Spatio-Temporal Video Grounding Task To ensure fairness and better highlight the effectiveness of our approach, we conduct instruction tuning on two powerful MLLMs, Qwen2.5-VL-7B and LLaVA-Video-7B, using the same UniViT dataset as *SpaceVLLM*, resulting in Qwen2.5-VL-SFT and LLaVA-Video-SFT.

For evaluation, following (Gu et al. 2024), we conduct STVG experiments on three benchmarks: HCSTVG-v1 (Tang et al. 2021), HCSTVG-v2(Tang et al. 2021), and VidSTG (Zhang et al. 2020). We use m_tIoU, m_vIoU, and vIoU@R as evaluation metrics, where m_tIoU measures temporal localization performance, while m_vIoU and vIoU@R evaluate spatial localization.

(1) HCSTVG-v1 and HCSTVG-v2. Table 2 presents the results on the HCSTVG-v1 test set, where our proposed method achieves state-of-the-art performance across all metrics. Compared to other MLLMs, *SpaceVLLM* attains a 39.5 m_vIoU score, outperforming Qwen2.5-VL-SFT by 10.9 points, and achieves 57.0 m_tIoU, surpassing it by 3.5 points. Relative to the base model LLaVA-Video-SFT, *SpaceVLLM* demonstrates improvements of 11.8 and 4.2 points in m_vIoU and m_tIoU, respectively. On the more comprehensive validation set of HCSTVG-v2, our method also achieves outstanding performance on all four metrics, as shown in Table 3. Specifically, *SpaceVLLM* outperforms the best-performing Qwen2.5-VL-SFT by 7.8 points in m_vIoU and by 3.4 points in m_tIoU. These results demonstrate that the decoupled head design enables MLLMs to generate more precise coordinates while also improving their temporal prediction capability.

(2) VidSTG. We evaluate the performance of *SpaceVLLM* on the more challenging VidSTG datasets in Table 4. Unlike HCSTVG's declarative-only annotation, the text captions in VidSTG include both declarative and interrogative sentences. In addition to our own trained baselines, we also

Models	m_tIoU	m_vIoU	vIoU@0.3	vIoU@0.5
Non-generative and task-specific models				
STGVT	-	18.2	26.8	9.5
TubeDETR	43.7	32.4	49.8	23.5
STVGFormer	-	36.9	**62.2**	34.8
CG-STVG	**52.8**	**38.4**	61.5	**36.3**
MLLMs with Parameter Sizes of 7B				
Qwen2.5-VL-SFT	53.5	28.6	45.2	21.9
LLaVA-Video-SFT	52.8	27.7	43.1	21.3
SpaceVLLM	**57.0**	**39.5**	**66.8**	**36.4**

Table 2: Comparison on HCSTVG-v1 test set (%).

Models	m_tIoU	m_vIoU	vIoU@0.3	vIoU@0.5
Non-generative and task-specific models				
MMN	-	30.3	49.0	25.6
TubeDETR	-	36.4	58.8	30.6
STVGFormer	58.1	38.7	**65.5**	33.8
CG-STVG	**60.0**	**39.5**	64.5	**36.3**
MLLMs with Parameter Sizes of 7B				
Qwen2.5-VL-SFT	55.3	26.5	38.6	20.2
LLaVA-Video-SFT	54.2	24.8	40.1	15.5
SpaceVLLM	**58.7**	**34.3**	**57.0**	**26.1**

Table 3: Comparison on HCSTVG-v2 val. set (%).

Models	Declarative Sentences				Interrogative Sentences			
	m_tIoU	m_vIoU	vIoU@0.3	vIoU@0.5	m_tIoU	m_vIoU	vIoU@0.3	vIoU@0.5
Non-generative and task-specific models								
STGVT (Tang et al. 2021)	-	21.6	29.8	18.9	-	-	-	-
TubeDETR (Yang et al. 2022)	48.1	30.4	42.5	28.2	46.9	25.7	35.7	23.2
STCAT (Jin et al. 2022)	50.8	33.1	46.2	32.6	49.7	28.2	39.2	26.6
STVGFormer (Lin et al. 2023b)	-	33.7	47.2	32.8	-	28.5	39.9	26.2
CG-STVG (Gu et al. 2024)	**51.4**	**34.0**	**47.7**	**33.1**	**49.9**	**29.0**	**40.5**	**27.5**
MLLMs with Parameter Sizes of 7B								
Qwen2.5-VL-SFT	41.6	20.3	26.1	15.4	40.9	17.1	17.6	13.9
LLaVA-Video-SFT	39.5	19.2	20.3	13.8	38.6	15.7	15.3	10.4
LLaVA-ST (Li et al. 2025)	45.5	24.8	36.0	22.9	43.2	20.0	28.1	17.5
SpaceVLLM	**48.7**	**27.5**	**40.5**	**25.5**	**48.8**	**25.2**	**36.7**	**21.6**

Table 4: Comparison with existing state-of-the-art models on VidSTG test set (%).

compare with LLaVA-ST, which adopts a more complex architecture and training strategy. The results demonstrate that *SpaceVLLM* outperforms all baselines, including the previous best model LLaVA-ST, across all metrics. Specifically, on declarative sentences, *SpaceVLLM* surpasses LLaVA-ST by 2.7 points in m_vIoU and 3.2 points in m_tIoU. On the more challenging interrogative sentences, it achieves improvements of 5.2 and 6.6 points in m_vIoU and m_tIoU, respectively. These results highlight that unlocking the inherent spatio-temporal understanding of MLLMs is more effective than imposing complicated external modules.

(3) Comparison with Task-specific Models. Notably, *SpaceVLLM* outperforms state-of-the-art task-specific models on HCSTVG-v1 and demonstrates remarkable generalization capabilities. Although it still falls short of task-specific models on certain metrics, it is important to emphasize that these specialized models are meticulously engineered for single tasks and require separate, resource-intensive training for each benchmark. In contrast, *SpaceVLLM* serves as a unified generative model that achieves strong performance across multiple tasks, including STVG, REC, and VTG with a single round of training.

Referring Expression Comprehension Task Our method is effective for spatio-temporal video grounding while also being seamlessly adaptable to image grounding tasks. Specifically, we simply introduce an image-specific special token, "<q_img>", which is trained to capture only spatial information. As shown in Table 5, on datasets such as RefCOCO (Kazemzadeh et al. 2014), RefCOCO+ (Kazemzadeh et al. 2014), and RefCOCOg (Mao et al. 2016), *SpaceVLLM* outperforms the previous best MLLMs (e.g., LLaVA-ST). For instance, it achieves a 2.1% improvement on the RefCOCO validation set and a 1% increase on the RefCOCO+ test-B set.

Video Temporal Grounding Task We compare our model with state-of-the-art models fine-tuned on Charades-STA (Gao et al. 2017) for video temporal grounding. As shown in Table 6, our approach achieves competitive performance compared to previously best-performing models. Furthermore, *SpaceVLLM* outperforms the baseline LLaVA-Video-SFT by 2.4 and 3.5 points in R1@0.5 and R1@0.7, respectively. These results demonstrate that the STA queries enhance the temporal sensitivity of the MLLM.

Video Understanding Task We further evaluate *SpaceVLLM* on multiple video understanding benchmarks, including MVBench (Li et al. 2024a), Egoschema (Mangalam, Akshulakov, and Malik 2023), TempCompass (Liu et al. 2024b), and VideoMME (Fu et al. 2024). As shown in Table 7, Our model supports various fine-grained grounding tasks while maintaining the general performance of the LLaVA-Video base model. Notably, *SpaceVLLM* achieves a significant 6.9% improvement on MVBench, particularly on spatio-temporal sub-tasks. This demonstrates that guiding the MLLM with STA queries enables it to better capture spatio-temporal details, thereby enhancing its performance on general spatio-temporal tasks.

Models	RefCOCO		RefCOCO+		RefCOCOg	
	val	test-A	val	test-A	val-u	test-u
Shikra-7B	87.0	90.6	81.6	87.4	82.3	82.2
GroundingGPT-7B	88.0	91.6	81.6	87.2	81.7	82.0
MiniGPT-v2-7B	88.7	91.7	80.0	85.1	84.4	84.7
Groma-7B	89.5	92.1	83.9	88.9	86.3	87.0
Qwen2.5-VL-7B	90.0	92.5	84.2	89.1	**87.2**	87.2
LLaVA-ST-7B	90.1	93.2	86.0	**91.3**	86.7	87.4
SpaceVLLM-7B	**90.7**	**93.4**	**86.3**	91.0	86.8	**88.1**

Table 5: Comparison on RefCOCO, RefCOCO+, and RefCOCOg (%). The accuracy with IoU threshold is 0.5.

Models	Type	R1@0.5	R1@0.7
SnAG (Mu, Mo, and Li 2024)	Trad.	64.6	46.2
EaTR (Jang et al. 2023)	Trad.	68.4	44.9
VTG-LLM-7B (Guo et al. 2024a)	MLLM	57.2	33.4
TRACE-7B (Guo et al. 2024b)	MLLM	61.7	41.4
TimeSuite-7B (Zeng et al. 2024)	MLLM	67.1	43.0
LLaVA-Video-SFT	MLLM	62.1	38.4
SpaceVLLM-7B	MLLM	64.5	41.9

Table 6: Comparison on Charades-STA for VTG in fine-tune settings. "Trad." refers to traditional models.

Ablation Study

In this section, we first compare and analyze the STVG performance of the architectures mentioned. Next, we conduct experiments to verify the effectiveness of the synthesized data. Finally, we compare the inference speed of our approach with other methods on the STVG task.

Model Architecture As shown in Table 8, we compare three decoder variants across STVG benchmarks, all trained on the UniViT. Among the three architectures, the simplest space head achieves the best overall performance, which supports our design principle. For the historical-assisted decoder, we set $n = 1$ and observe that in the later stages of training, the gating weight $\alpha_{i,i}$ for $\hat{\mathbf{Q}}_i$ is consistently equal to or close to 1. This indicates that the queries have already fully interacted with the historical frames within the MLLM without relying on additional historical queries. Notably, the most complex DETR-like architecture yields the lowest accuracy. This demonstrates that the STA queries encoded by the LLM are already highly refined and interpretable, allowing downstream prediction to achieve promising results with a lightweight head. In contrast, a complex architecture may introduce noise or disrupt the existing representations.

Dataset To verify the effectiveness of our proposed dataset V-STG, we conduct ablation studies as shown in Table 9. We find that incorporating our synthetic data leads to improvements of 2.7% in m_tIoU and 5.1% in m_vIoU. It demonstrates that our synthetic data is effective for improving the MLLM's spatio-temporal video grounding capability.

Inference Efficiency Inference speed for the STVG task is another advantage of our method. Previous approaches generate spatial coordinates for each frame sequentially, resulting in a long, token-by-token output process. In contrast, *SpaceVLLM* batches STA queries and feeds them into the space head to output all required coordinates simultaneously. On the VidSTG declarative test set (4609 instances) using a single H100 GPU without any acceleration, LLaVA-ST (Li et al. 2025) takes 17.63 seconds per sample on average, while *SpaceVLLM* only takes 2.89 seconds (Table 10).

Models	MVBench	EgoSch.	TempCom.	VideoMME (w-subs)
LLaVA-Video-7B	58.6	57.3	67.0	63.3 / 69.7
LLaVA-Video-SFT	61.6	57.7	67.1	62.0 / 69.9
SpaceVLLM-7B	65.5	57.4	67.3	62.6 / 70.4

Table 7: Comparison on MVBench, EgoSchema, TempCompass, and VideoMME for video understanding task (%).

Methods	HCSTVG-v1	VidSTG-D
DETR-like Architecture	56.1 / 35.5	46.2 / 24.1
Historical-Assisted Decoder	55.5 / 38.5	47.5 / 26.6
Query-Guided Space Head	**57.0 / 39.5**	**48.7 / 27.5**

Table 8: Ablation studies on architecture of *SpaceVLLM*. VidSTG-D denotes declarative sentences. Results are reported as "m_tIoU / m_vIoU" for each benchmark.

Datasets	m_tIoU	m_vIoU	vIoU@0.3	vIoU@0.5
V-STG w/o synthetic data	54.3	34.4	57.6	25.4
V-STG	**57.0**	**39.5**	**66.8**	**36.4**

Table 9: Ablation studies on proposed dataset, evaluated on the HCSTVG-v1 test set.

Model	Total Time(s)	Avg Time per Sample (s)
LLaVA-ST	81256	17.63
SpaceVLLM	13323	2.89

Table 10: Inference speed comparison of different models.

Conclusion

In this paper, we introduce *SpaceVLLM*, a MLLM with spatio-temporal video grounding (STVG) capability. Our design principle is unlocking the inherent potential of MLLMs and leveraging their strengths. Specifically, we employ Spatio-Temporal Aware Query to guide the MLLM in extracting spatio-temporal information. We then utilize a lightweight Query-Guided Space Head, which relieves the MLLM from the burden of spatial coordinate generation. Moreover, we introduce a Video Spatio-Temporal Grounding (V-STG) dataset to advance multimodal spatio-temporal understanding. Extensive experiments demonstrate that our model achieves state-of-the-art performance on multiple benchmarks, fully validating the effectiveness of our model.

Acknowledgments

This work is supported by the National Nature Science Foundation of China (62425114, 62121002, U23B2028, 62232006). We acknowledge the support of the GPU cluster built by MCC Lab of Information Science and Technology Institution, USTC, and USTC super-computing center for providing computational resources for this project.

References

Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; et al. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Chen, K.; Zhang, Z.; Zeng, W.; Zhang, R.; Zhu, F.; and Zhao, R. 2023. Shikra: Unleashing multimodal llm's referential dialogue magic. *arXiv preprint arXiv:2306.15195*.

Chen, L.; Wei, X.; Li, J.; Dong, X.; Zhang, P.; Zang, Y.; Chen, Z.; Duan, H.; Lin, B.; Tang, Z.; et al. 2024. Sharegpt4video: Improving video understanding and generation with better captions. *arXiv preprint arXiv:2406.04325*.

Fu, C.; Dai, Y.; Luo, Y.; Li, L.; Ren, S.; Zhang, R.; Wang, Z.; Zhou, C.; Shen, Y.; Zhang, M.; et al. 2024. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. *arXiv preprint arXiv:2405.21075*.

Gao, J.; Sun, C.; Yang, Z.; and Nevatia, R. 2017. Tall: Temporal activity localization via language query. In *Proceedings of the IEEE International Conference on Computer Vision*, 5267–5275.

Gu, X.; Fan, H.; Huang, Y.; Luo, T.; and Zhang, L. 2024. Context-guided spatio-temporal video grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18330–18339.

Guo, Y.; Liu, J.; Li, M.; Cheng, D.; Tang, X.; Sui, D.; Liu, Q.; Chen, X.; and Zhao, K. 2024a. VTG-LLM:Integrating Timestamp Knowledge into Video LLMs for Enhanced Video Temporal Grounding. *arXiv preprint arXiv:2405.13382*.

Guo, Y.; Liu, J.; Li, M.; Cheng, D.; Tang, X.; Sui, D.; Liu, Q.; Chen, X.; and Zhao, K. 2025. Vtg-llm: Integrating timestamp knowledge into video llms for enhanced video temporal grounding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 3302–3310.

Guo, Y.; Liu, J.; Li, M.; Tang, X.; Liu, Q.; and Chen, X. 2024b. Trace: Temporal grounding video llm via causal event modeling. *arXiv preprint arXiv:2410.05643*.

Hendricks, L. A.; Wang, O.; Shechtman, E.; Sivic, J.; Darrell, T.; and Russell, B. 2017. Localizing moments in video with natural language. In *Proceedings of the IEEE International Conference on Computer Vision*, 5803–5812.

Huang, B.; Wang, X.; Chen, H.; Song, Z.; and Zhu, W. 2024. Vtimellm: Empower llm to grasp video moments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14271–14280.

Jang, J.; Park, J.; Kim, J.; Kwon, H.; and Sohn, K. 2023. Knowing where to focus: Event-aware transformer for video grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 13846–13856.

Jin, Y.; yongzhi li; Yuan, Z.; and Mu, Y. 2022. Embracing Consistency: A One-Stage Approach for Spatio-Temporal Video Grounding. In *Advances in Neural Information Processing Systems*, 29192–29204.

Kazemzadeh, S.; Ordonez, V.; Matten, M.; and Berg, T. L. 2014. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing*, 787–798.

Li, H.; Chen, J.; Wei, Z.; Huang, S.; Hui, T.; Gao, J.; Wei, X.; and Liu, S. 2025. Llava-st: A multimodal large language model for fine-grained spatial-temporal understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 8592–8603.

Li, K.; He, Y.; Wang, Y.; Li, Y.; Wang, W.; Luo, P.; Wang, Y.; Wang, L.; and Qiao, Y. 2023. Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*.

Li, K.; Wang, Y.; He, Y.; Li, Y.; Wang, Y.; Liu, Y.; Wang, Z.; Xu, J.; Chen, G.; Luo, P.; et al. 2024a. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22195–22206.

Li, Z.; Xu, Q.; Zhang, D.; Song, H.; Cai, Y.; Qi, Q.; Zhou, R.; Pan, J.; Li, Z.; Vu, V. T.; Huang, Z.; and Wang, T. 2024b. Groundinggpt: Language enhanced multi-modal grounding model. *arXiv preprint arXiv:2401.06071*.

Lin, B.; Ye, Y.; Zhu, B.; Cui, J.; Ning, M.; Jin, P.; and Yuan, L. 2023a. Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*.

Lin, Z.; Tan, C.; Hu, J.-F.; Jin, Z.; Ye, T.; and Zheng, W.-S. 2023b. Collaborative static and dynamic vision-language streams for spatio-temporal video grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 23100–23109.

Liu, S.; Zeng, Z.; Ren, T.; Li, F.; Zhang, H.; Yang, J.; Jiang, Q.; Li, C.; Yang, J.; Su, H.; Zhu, J.; and Zhang, L. 2024a. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European Conference on Computer Vision*, 38–55.

Liu, Y.; Li, S.; Liu, Y.; Wang, Y.; Ren, S.; Li, L.; Chen, S.; Sun, X.; and Hou, L. 2024b. Tempcompass: Do video llms really understand videos? *arXiv preprint arXiv:2403.00476*.

Loshchilov, I.; and Hutter, F. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Ma, C.; Jiang, Y.; Wu, J.; Yuan, Z.; and Qi, X. 2024. Groma: Localized Visual Tokenization for Grounding Multimodal Large Language Models. In *European Conference on Computer Vision*, 417–435.

Maaz, M.; Rasheed, H.; Khan, S.; and Khan, F. 2024a. Videogpt+: Integrating image and video encoders for enhanced video understanding. *arXiv preprint arXiv:2406.09418*.

Maaz, M.; Rasheed, H.; Khan, S.; and Khan, F. S. 2024b. 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*, 12585–12602.

Mangalam, K.; Akshulakov, R.; and Malik, J. 2023. Egoschema: A diagnostic benchmark for very long-form video language understanding. *Advances in Neural Information Processing Systems*, 36: 46212–46244.

Mao, J.; Huang, J.; Toshev, A.; Camburu, O.; Yuille, A. L.; and Murphy, K. 2016. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 11–20.

Mu, F.; Mo, S.; and Li, Y. 2024. Snag: Scalable and accurate video grounding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 18930–18940.

Patraucean, V.; Smaira, L.; Gupta, A.; Recasens, A.; Markeeva, L.; Banarse, D.; Koppula, S.; Malinowski, M.; Yang, Y.; Doersch, C.; et al. 2023. Perception test: A diagnostic benchmark for multimodal video models. *Advances in Neural Information Processing Systems*, 36: 42748–42761.

Regneri, M.; Rohrbach, M.; Wetzel, D.; Thater, S.; Schiele, B.; and Pinkal, M. 2013. Grounding action descriptions in videos. In *Transactions of the Association for Computational Linguistics*, 25–36.

Rezatofighi, H.; Tsoi, N.; Gwak, J.; Sadeghian, A.; Reid, I.; and Savarese, S. 2019. Generalized Intersection over Union: A Metric and A Loss for Bounding Box Regression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 658–666.

Share. 2024. ShareGemini: Scaling Up Video Caption Data for Multimodal Large Language Models. https://github.com/Share14/ShareGemini.

Tang, Z.; Liao, Y.; Liu, S.; Li, G.; Jin, X.; Jiang, H.; Yu, Q.; and Xu, D. 2021. Human-centric spatio-temporal video grounding with visual transformers. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(12): 8238–8249.

Wang, Y.; He, Y.; Li, Y.; Li, K.; Yu, J.; Ma, X.; Li, X.; Chen, G.; Chen, X.; Wang, Y.; He, C.; Luo, P.; Liu, Z.; Wang, Y.; Wang, L.; and Qiao, Y. 2023. Internvid: A large-scale video-text dataset for multimodal understanding and generation. In *arXiv preprint arXiv:2307.06942*.

Wu, B.; Yu, S.; Chen, Z.; Tenenbaum, J. B.; and Gan, C. 2024. Star: A benchmark for situated reasoning in real-world videos. *arXiv preprint arXiv:2405.09711*.

Xiao, J.; Shang, X.; Yao, A.; and Chua, T.-S. 2021. Next-qa: Next phase of question-answering to explaining temporal actions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9777–9786.

Yang, A.; Miech, A.; Sivic, J.; Laptev, I.; and Schmid, C. 2022. Tubedetr: Spatio-temporal video grounding with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16442–16453.

Yang, A.; Yang, B.; Hui, B.; Zheng, B.; Yu, B.; Zhou, C.; Li, C.; Li, C.; Liu, D.; Huang, F.; et al. 2024a. Qwen2 Technical Report. *arXiv preprint arXiv:2407.10671*.

Yang, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Li, C.; Liu, D.; Huang, F.; Wei, H.; et al. 2024b. Qwen2.5 Technical Report. *arXiv preprint arXiv:2412.15115*.

Yi, K.; Gan, C.; Li, Y.; Kohli, P.; Wu, J.; Torralba, A.; and B.Tenenbaum, J. 2020. Clevrer: Collision events for video representation and reasoning. In *International Conference on Learning Representations*.

Yu, Z.; Xu, D.; Yu, J.; Yu, T.; Zhao, Z.; Zhuang, Y.; and Tao, D. 2019. Activitynet-qa: A dataset for understanding complex web videos via question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 9127–9134.

Zala, A.; Cho, J.; Kottur, S.; Chen, X.; Oguz, B.; Mehdad, Y.; and Bansal, M. 2023. Hierarchical video-moment retrieval and step-captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 23056–23065.

Zeng, X.; Li, K.; Wang, C.; Li, X.; Jiang, T.; Yan, Z.; Li, S.; Shi, Y.; Yue, Z.; Wang, Y.; et al. 2024. Timesuite: Improving mllms for long video understanding via grounded tuning. *arXiv preprint arXiv:2410.19702*.

Zhai, X.; Mustafa, B.; Kolesnikov, A.; and Beyer, L. 2023. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, 11975–11986.

Zhang, H.; Li, X.; and Bing, L. 2023. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*.

Zhang, Y.; Wu, J.; Li, W.; Li, B.; Ma, Z.; Liu, Z.; and Li, C. 2024. Video instruction tuning with synthetic data. *arXiv preprint arXiv:2410.02713*.

Zhang, Z.; Zhao, Z.; Lin, Z.; Huai, B.; and Yuan, J. 2021. Object-aware multi-branch relation networks for spatio-temporal video grounding. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, 1069–1075.

Zhang, Z.; Zhao, Z.; Zhao, Y.; Wang, Q.; Liu, H.; and Gao, L. 2020. Where does it exist: Spatio-temporal video grounding for multi-form sentences. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, 10668–10677.